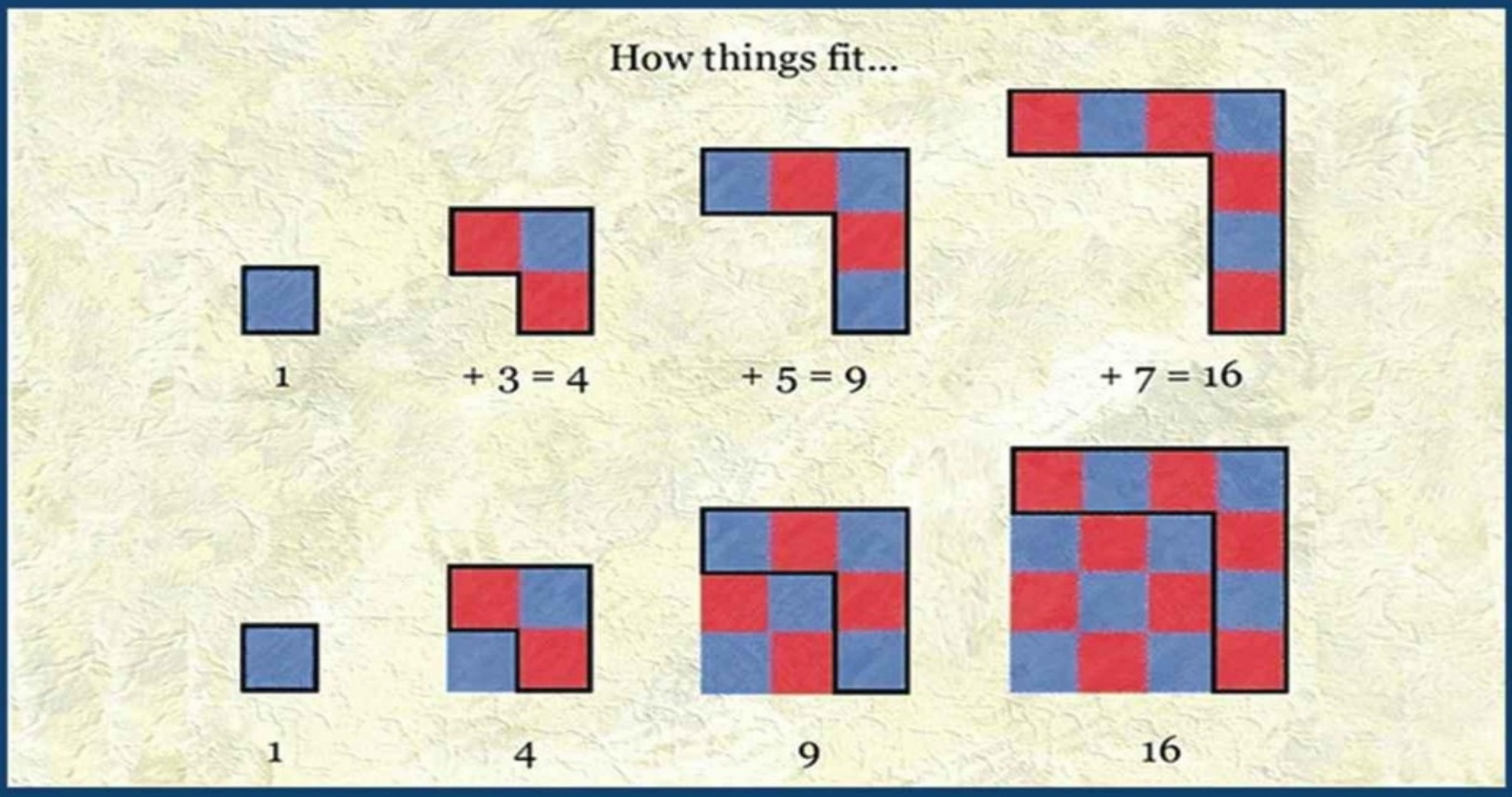
Mathematics is About the World

How Ayn Rand's theory of concepts unlocks the false alternatives between Plato's Mathematical Universe and Hilbert's Game of Symbols



Robert E. Knapp

MATHEMATICS
IS ABOUT THE WORLD

MATHEMATICS
IS ABOUT THE WORLD

HOW AYN RAND'S THEORY OF CONCEPTS UNLOCKS THE FALSE ALTERNATIVES BETWEEN PLATO'S MATHEMATICAL UNIVERSE AND HILBERT'S GAME OF SYMBOLS

Robert E. Knapp

First Edition

Visit the website for the book: www.MathematicsIsAboutTheWorld.com to find out more about the book and to contact the author.

CONTENTS

Acknowledgements

My greatest debt is to Ayn Rand's *Introduction to Objectivist Epiestemology* without which this book could not have been conceived. Published during my college years, it has illuminated my understanding and appreciation for mathematics across five decades.

I have been stimulated by a number of extended exchanges with Harry Binswanger on a variety of mathematical topics. In regards to my own work, Harry has provided welcome encouragement and valuable ciriticism that have helped me sharpen my formulations and, indeed, saved me from several serious mistakes.

Pat Corvini's fascinating and stimulating lectures on Zeno's paradox, together with the reactions of its audience first convinced me that there was a potential audience for my own ideas and triggered my first essays on how to look at mathematics.

Objectivist Conferences (OCON) provided my first opportunity for a public presentation of my ideas and I am grateful for the questions and comments from my audience, and especially to Ray and Rebecca Girn, Evan Picoult, and Irene Knapp who previewed my courses and offered many valuable comments and suggestions.

I have received enouragement and valuable comments on chapters and/or earlier formulations of these ideas from Harry Binswanger, Raymond Knapp, Evan Picoult, Ray Girn, Rebecca Girn, Robert Rubinstein, Kenneth and Sandy Landin, Judy Knapp, and Shrikant Rangnekar.

Shrikant was particularly helpful in the earlier stages of this book, reading and critiquing every chapter as it was written, and making a second careful reading of the initial draft. His helpful observations and suggestions, interest, and encouragement was at once helpful and very motivating. His guidance through my early attempts to find a publisher or an agent, though unsuccessful, netted much of the material that has made it into my preface. Throughout, his knowledge and judgement of the publishing industry has been invaluable and his enthusiasm for my book has been inspiring.

Finally, it should go without saying that all responsibility for both the content and presentation of that content is entirely my own.

Preface

What is mathematics about? Is there a world of mathematics in which straight lines are infinitely straight, are infinitely thin, and contain an uncountable infinity of points? Or, on the contrary, is mathematics a purely formal system of

deductions from axioms, a system that is beautiful to some of us, but boring and just too difficult for the rest of us? Is it devoid, in either case, of any worldly or other worldly content, yet

mysteriously applicable to the world we inhabit? Or, finally, contrary to the nearly universal range of acceptable academic views, is mathematics about the world?

This book is written for people who care about such questions, people who are still looking for answers and are not satisfied with the standard suggestions, people who are open to a common sense-yet radical-alternative.

This is the book I would have loved to find when I encountered abstract mathematics in my high school years and later, in college, began the process of becoming a mathematician.

It's the alternative perspective I needed when I read, also in high school, Russell's *Introduction to Mathematical Philosophy.* Though from a far different perspective, like Russell's, mine is a book on how to think about mathematics.

My own fascination with mathematics began with a book by Irving Adler entitled *Magic House of Numbers*, which I checked out of the library in sixth grade because of the word "magic" in the title. Thus began a lifetime of interest, study, and for a brief time research, in mathematics. By the end of ninth grade, with the aid of old textbooks lying around the house, I was learning trigonometry, analytic geometry, and the beginnings of calculus.

In my sophomore year of high school, I found a copy of Birkoff and MacLane's highly regarded *A Survey of Modern Algebra* in my school library. And I was struck by the opening of Chapter III, "Polynomials", which reads "Let D be any integral domain, and let "x" be any symbol. Suppose one forms sums,

products, and differences of x with the elements of D and with itself, subject to the rules of ordinary algebra…"

Now it is true that a polynomial defines a rule that can be applied to any mathematical domain that employs a sensible notion of addition and multiplication. One can, for example, apply

polynomials to matrices and even to other polynomials. So a

polynomial has a meaning and significance as a set of rules that transcends any particular mathematical domain to which it might be applied.

But this is not what I took the authors to mean, nor what they were actually saying. Instead, I heard: Don't think of

polynomials as meaning anything in particular. Look at them as manipulations of meaningless symbols. Well, I continued for many more pages before I returned the book to the library, but I got nothing out of it. Such was my first encounter with abstract mathematics!

By the end of my junior high school year I had read about matrices and vector spaces from an interesting perspective that made sense to me. Then, at some point near the end of that year or, perhaps, that summer I started learning about "rings" and "ideals"

as an abstract study. I had no idea why "ideals" were important or what they were about. That "rings" were a general framework for studying divisibility and factorization never occurred to me. But I remember thinking, "Rings are just a way of thinking about

numbers, matrices, and polynomials." From that moment, I

embraced mathematical abstraction. Yet, simultaneously, I

embraced the view that mathematics is about the world. I had glimpsed that mathematical abstractions are a way of looking at things; neither a path to a separate mathematical world nor a play with meaningless symbols.

During my senior year in high school I started reading, with great appreciation, Halmos' great text on finite dimensional vector spaces, presented on an abstract

level, but very readable and inspiring. Later, I began, again with appreciation, my freshman year in college reading Dieudonne's highly abstract *Foundations of Modern Analysis.* Both books reinforced my belief in the power of mathematical abstraction; neither deterred me from my view that mathematics is about the world.

At some point during my college freshman year, I realized that neither mathematicians nor philosophers of mathematics

shared my perspective, offering only the alternatives of formalism (a game of symbol manipulation), Platonism (a separate world of mathematics), or, as a third, the Fregean view that mathematics is essentially a branch of logic. I could accept none of these choices.

However, I had discovered Ayn Rand and during the following summer she began her epochal series of articles on Objectivist epistemology. It struck me, from her first essay linking mathematics and concept formation, that I had found my key to understanding mathematics. In regards to its specific application to mathematics, however, I also realized that I was on my own.

The purpose of this book is to share my present understanding and perspective, to offer a distinctive view of mathematics that was made possible for me by Ayn Rand's essays and extemporaneous comments relating to her theory of universals.

And my central message is this: You do not have to choose between mathematical abstractions and reality. Mathematical abstractions are a way of understanding the world, of deepening and enriching one's perspective on the world. One understands the essence of a mathematical discipline when one grasps what it is trying to measure. Mathematics is about the world.

Though I accept neither alternative, I find an important difference between the ways that mathematicians and philosophers each typically look at mathematics. Mathematicians in general, operating in a world of "complex manifolds", "homological

algebra", infinite-dimensional "function spaces", and "fiber bundles" tend to believe in some kind of Platonic mathematical world suggested by, but distinct from, the world we live in and taking on a life of its own. They know that mathematics is hard work and must involve something more than arbitrary

symbols and rules. Philosophers, on the other hand, generally dismiss a world of mathematics. They differ on what should take its place. Some of the best, certainly, have attempted a naturalistic, reality-based, approach to understanding mathematics, starting with ordinary numerical concepts of the sort we form in childhood. But they typically try to incorporate formal set theory as it was developed during the 20th century and consequently, at some point along the way, they leave the world behind.

But is there a third alternative? When I say to someone who is neither a mathematician nor a philosopher, “Mathematics is about the world,” the usual response is, “Of course!” But such responses quickly change when I bring up infinitely straight infinitely thin straight lines and, in general, mathematical infinity.

How can there be an infinite number of counting numbers? Or, on the contrary, is there a finite number? Is there a point, a large number “N” such that N is meaningful, but N + 1 is meaningless?

And, if so, what about the immediate implication that the equation x – 1 = B has a solution for B = 5 or for B = 500 trillion, but does not have a solution for this “largest” number B = N? Or what about the fact that, for this largest number I am calling N, we would not be able to express a length of N feet in inches, because we would have to multiply N times 12 to find the number of inches resulting in a number larger than N?

It appears that a number system, to be manageable, must be open-ended. Mathematical infinities that involve numbers, infinities that do not exist in the world,

are,

indeed,

an

indispensable part of mathematics. Similarly, in geometry, the infinite precision of Euclid’s geometric reasoning is necessary to the entire structure and the relationships that Euclid uncovers must be utilized in any attempt to measure degrees of imprecision in one’s measurements of objects in the world.

In sum, it seems that the more one thinks about mathematics and its methods (and the more advanced one's

mathematical studies) the harder it becomes to maintain, as I nonetheless do, that mathematics is about the world. These

paradoxes are obvious and must be answered: If mathematical

infinities do not exist in the world, how can a mathematics that includes mathematical infinities be about the world? Or, if all geometric measurements have finite precision, yet geometric

concepts and arguments are infinitely precise, how can geometry be about the world?

So this book is, necessarily, a defense of an unpopular viewpoint; but more fundamentally it is a book about how to think about mathematics.

As a pursuit, mathematics is about discovering relationships between quantities, about discovering and solving algebraic and differential equations or finding geometric

relationships. It is, as Ayn Rand put it, the science of measurement.

But why is measurement important? And why do the demands of

measurement require the amazingly abstract complex science that mathematics became during the 19th and 20th centuries? Do all these abstractions really have something to do with measurement?

I offer my affirmative answer, in part, as the key to relating mathematics to the world. But I also believe that the measurement viewpoint is the key to understanding the underlying purpose and the logical structure of mathematics.

In this book, I apply my viewpoint, first to elementary and then to more advanced mathematical topics. I start with elementary topics to show how mathematics begins with measurement and to show how mathematical progress is driven by the needs of

measurement. I proceed to more advanced topics to indicate, by selected examples, how measurement remains the underlying

thread as one ascends to higher and higher degrees of mathematical abstraction.

So, why is measurement important?
It is measurement that enables us to make fine distinctions,

to specify differences among similar shared characteristics.

Measurement comes into play whenever we weigh alternatives. It shows up in daily life when we ask: How far is it? How fast am I going? How long will it take me to get there? Can I afford to buy that new car? How much will it cost? What will I have to give up or cut back on? What size of refrigerator will fit in that space? How high can it be? How wide? What will it hold? How much do I

weigh? How much weight should I try to lose?
Such questions occur whenever we want to judge

differences,

consider alternatives, and specify

objectives.

Measurements are a way of determining precisely what already is and of specifying precisely what you want to bring into existence.

Measurements specify relationships, quantitative or causal, that exist or might exist in the world.
But measurement applies even more broadly to the other

sciences. When causal quantitative relationships are identified in general form, as in physical laws of nature, they apply generically to a broad range of cases. Based on Newton's laws of motion, if I shoot a cannon ball at a known angle and known velocity, I can calculate mathematically how high it will rise and where it will fall. In fact, I can compute, from the starting angle and landing point, its entire trajectory! Newton's laws of motion, originally discovered by means of measurement, become, in turn, the means of further

measurements. Measurement is one key to understanding the

world and adapting it to our own purposes.
But, given that measurement is important, why do we need

But, given that measurement is important, *why* do we need

a vast, complex, abstract science to help us with our

measurements?
If I want to measure my couch, I take out a tape measure. If

I want to find out what I weigh, I step on a scale. If I want to know whether I'm speeding, I check the speedometer. But how do I

measure the circumference of the earth or the distance to the sun or the distance to the moon? How do we discover the mass of Jupiter or the average speed of Mercury as it orbits the sun? All of these determinations involve measurement, but in none of these cases do we simply measure, directly, the quantity we are trying to specify.
Yet, amazingly, the circumference of the earth

was

measured (within about 16%) by Eratosthenes in 200 BC, without leaving Alexandria. He did it by measuring the angle the sun's rays made at noon on the day of the summer solstice. He made just this one measurement, but relied, for his calculation on two others: first, the distance of a particular town to his south and, second, the known fact that, in that town, at that time, and on that day of the summer solstice, the sun was directly overhead, a circumstance manifested by the known observation that the one could see, at the stroke of noon, the reflection of the sun at the bottom of a very deep well. And what made this possible? His knowledge of Euclidean geometry!
But this is just a dramatic example of a universal pattern.

Most of the measurements we make are indirect in one form or another. When we step on a scale, we rely on a hidden mechanism, calibrated based on known physical laws that required mathematics in both their discovery and their application. Every gauge, every speedometer, every mechanical or electrical measuring device utilizes indirect measurement to determine the quantitative

relationship one is trying to ascertain.
Whether one measures a shadow to find the height of a

flagpole or solves a differential equation to discover the trajectory of a projectile, one is relying on direct measurements of one set of quantities, in the context of

previously discovered relationships, to determine, indirectly, the desired measurements of a second set of quantities. This is indirect measurement. It is the mathematical relationships and the physical laws discovered by their use that make indirect measurement possible. And it is the need for indirect measurement that explains the need for a science of mathematics.
Progress in mathematics consists in finding the

connections, in finding the geometric and mathematical

relationships that make indirect measurement possible. Every geometric theorem, every algebraic or differential equation

expresses a relationship that can provide a bridge to an indirect measurement. But an equation is also something one needs to solve.

Whenever certain variables of an equation are regarded as known, while others are regarded as unknown, the challenge is to “solve”

the equation, to discover the corresponding values of the unknown variables. And this challenge, the search for solutions and for general methods of finding solutions, has driven progress in mathematics. Every mathematical abstraction and every new

mathematical relationship provides one more step on this complex journey. In general, every part of mathematics, from the most elementary to the most abstract, began with a problem in indirect measurement and can be better understood in relation to

measurement. Indirect measurement is the heart of mathematics, its reason for being, and the source of its power to enrich our lives.
I apply the measurement perspective throughout the book,

providing a range of examples that illustrate its applicability to both elementary and more advanced mathematics.
I have written this book for a general audience. And I have

endeavored to make it accessible and interesting to philosophers, mathematicians, advanced high school students, and interested laymen. To philosophers, I offer a serious, if unconventional, alternative to existing views; I offer a non-Platonic form of realism as a way to look at mathematics. To

mathematicians, I offer an account of just what sort of thing they are discovering and why these discoveries are important. I thereby aim to demystify the obvious applicability of mathematics, including advanced, abstract mathematics, to the world. To high school and college students interested in pursuing mathematics, I offer an integrating

perspective that will help them learn and appreciate the concepts and methods of advanced mathematics. And, to the educated

laymen, I offer a new way to think about mathematical concepts that will widen their perspective and illuminate their other readings in mathematics.
The book is organized into two parts. The first part,

consisting of five chapters, is elementary, should be accessible to high school students, and discusses plane Euclidean geometry and the real number system. The first two chapters are intended to be read first. One can read every chapter independently without getting lost, but the first two chapters provide the best context for the two that follow and, indeed, for the rest of the book.
The last three chapters address more advanced and more

modern topics, namely set theory, point set topology, linear algebra, and, from an introductory perspective, group

representations.
I conclude this introduction with a brief summary of every

chapter in the book:
Chapter 1 exhibits all of Euclid's postulates as primitive

measurements, while emphasizing that all measurements of

concretes are subject to specific precision requirements. Euclid's arguments are a form of indirect measurement, are, in essence, recipes for a series of measurements.
The first chapter addresses questions such as: How can

Euclid's arguments be rigorously valid if they pertain to realworld shapes and to realworld geometric relationships? More generally, given that all measurement is subject to finite precision limits, why, from a reality-based perspective, is it possible, meaningful, and necessary for mathematics to be infinitely precise?
Chapter 2 analyzes magnitude, geometrically, as an object

Chapter 2 analyzes magnitude, geometrically, as an object

of numerical measurement. Relationships between numbers reflect quantitative relationships in the world among the quantities that they measure.
Chapter 3 further examines the meaning and consequences

of Euclid's fifth postulate, following Euclid to show just how the measurement of area and the laws of geometric proportion both depend on the properties of parallel lines.
Chapter 4 shows how irrational numbers relate to the world

and why are they needed in mathematics. It identifies the realworld meaning of convergence and completeness of the real number system. It sorts out the important mathematical

contributions from the alleged philosophical implications of two celebrated constructions of the real number system,

one

by

Dedekind and the other by Cantor.
The fourth chapter addresses questions such as: Why are

irrational numbers needed in mathematics, despite being

indistinguishable, in any concrete application, from rational numbers? In this connection, what is the realworld meaning of convergence and of the completeness of the real number system?
Chapter 5 considers the

pervasive role of geometric

abstraction in mathematics, as it relates to measurement. It shows how geometry provides a conceptual perspective to think about actual objects and relationships in the world: objects and

relationships far beyond the classical confines of plane and solid geometry.
Set theory has become indispensable to modern

mathematics. Yet the contradictions in the original naïve

mathematics. Yet the contradictions in the original naive

conception of sets led to the ontologically meaningless ZermeloFraenkel axioms of set theory, as a purported foundation of mathematics. Chapter 6 provides an alternative realist account and rationale for the use of set theory in mathematics, emphasizing the importance of a proper hierarchy of mathematical abstraction. Chapter 6 addresses the question: What are the actual

need, the proper context, and the key function of set theory in mathematics? Chapter 7 explores the ways that the measurement

perspective illuminates and integrates our understanding of vector spaces and linear algebra.
Chapter 8 examines a realm that is often thought to have

little or no relationship to quantity, namely group theory. It shows how groups arise, why they are important, and, in just what sense the symmetry that they measure sits at the heart of the conceptual process itself.
This last chapter asks and answers the question: What do

abstract groups have to do with measurement?
Two themes run throughout the book: The first is that

mathematics is about the world. And the second is that indirect measurement is the heart of mathematics, its reason for being, and the source of its power to enrich our lives.

MATHEMATICS
IS ABOUT THE WORLD

PART 1: ELEMENTARY

Chapter 1

Euclid’s Method

It should not be controversial to observe that mathematics

is about the world. But it is.
Classical Greek mathematics did not end with Euclid. His

successor, Archimedes, is universally recognized as one of the

greatest mathematicians of all time. Even so, Euclid’s *Elements* was the culmination of all that preceded it and the foundation of all that

followed. The pinnacle of Greek mathematics, Euclid’s *Elements* remained unsurpassed for two millennia.
But the mathematical confusion begins with Euclid. “A

point,” says Euclid, “is that which has no part. A line is a

breadthless length.”[1] So begins Book I with definitions 1 and 2. And then Euclid proceeds to illustrate his propositions with points that

do have parts and lines that do have width, illustrations that,

accordingly, are visible to the human eye.
So what do his propositions mean? Do they pertain to

relationships in the world and to the pictures that Euclid draws for

us? Or, on the contrary, are Euclid’s pictures merely imperfect,

suggestive renderings of ideal geometric figures? And do such

geometric figures constitute a world of their own consisting of ideal

points that have no extension and of ideal lines that have no width?

The common fallacy that mathematics pertains specifically to a

mathematical universe; that mathematics *applies* to the world, but is not *about* the world begins with the first page of Euclid's monumental work.
I call it a fallacy, because I do not share it. Mathematics

applies to the world *because* it is about the world.
But this confusion is even older than Euclid and it persists

to this day. Today, it is widely held, by philosophers,

mathematicians, and educated laymen, that Euclid's propositions

are only true for idealized figures in a geometric universe. And that

his arguments only apply rigorously to objects in that universe.
This is Platonism and it dates back to Plato, who preceded

Euclid. In the lexicon of today's philosophers of mathematics and among mathematicians, it is the so-called "realist" view.[2]Certainly there are alternatives, but the dominant alternative to this realism

sees mathematics as purely formal. This competing view that

mathematics consists of deductions from arbitrary axioms, goes

back to David Hilbert in the late 19th and early 20th century.[3] In Willard

Quine's words, "… the formalist keeps classical mathematics as a play of insignificant notations."[4] Today the view that either geometry or mathematics in general could be about this

earth, about the world we inhabit, is generally regarded as untenable.[5]
As Plato puts it, "[geometers] make use of visible figures

and discourse about them though what they really have in mind is

the originals of which these figures are images."[6] What are these originals? Plato here refers to Platonic archetypical, idealized

figures residing in Plato's famous World of Forms or Ideas, as

figures, residing in Plato's famous world of Forms or Ideas, as

described in Plato's *Republic*. Generally, Plato held that objects in the world are imperfect reflections of the archetypes residing in his world of Forms.[7] And although Plato's world of Forms is no longer taken seriously, the general viewpoint, that geometric figures are

idealizations, did not die with the eclipse of Plato's heaven. John

Stuart Mill, for example, characterizes geometric figures as

something "painted in the imagination" to which one compares the shapes of objects or drawings in the world.[8] And in this comparison, as Plato had maintained, the world is found wanting.

Platonism is not considered a satisfactory view today. But

Penelope Maddy speaks for the consensus when she says, "Let me

return now to Platonism, the view that mathematics is an objective

science."[9] And how do modern philosophers, characterize Platonism? Stewart Shapiro summarizes Plato's view, as follows:

"some physical objects approximate Euclidean figures. But geometric theorems do not apply to these approximations."[10] And Philip Kitcher elaborates, "... the Platonist thesis: true

mathematical statements are true in virtue of the properties of abstract objects."[11]

The modern consensus, then, is that Platonism, in a

modern reincarnation, provides the only available viewpoint of

mathematics as an objective science. And, in search of such

objectivity, philosophers such as Kitcher, Maddy, Shapiro, Resnik,

and Parsons have made serious, sophisticated, attempts to develop

a third alternative, some kind of modified non-mystical Platonism.

However, my own viewpoint, that geometry is about the world, this world, has not been in the running, has been ruled out of court, has not been taken seriously.

The Paradox for a Proper Realism

So what is the difficulty?
The mathematical issue arises from the fact that all measurement of continuous quantities is approximate. We measure our own height in inches and, perhaps, achieve accuracy within a quarter of an inch. We measure our weight in pounds and the answer is usually accurate within half a pound. Within an appropriate context, more accurate measurements of length and weight are possible, but one never achieves infinite precision. There is always a limit to the subdivisions that one makes and to the accuracy with which one can apply them. All precision is finite.
But no acknowledgement of such precision limits can be found in Euclid. Euclid's equilateral triangles have sides of absolutely equal length and angles that are absolutely equal. When he bisects a line he treats the bisection as totally exact. Euclid's lines are totally straight and the points on his circles are all exactly the same distance from its center, a central point that has no parts.
In light of the mathematical issue, the philosophical issue

is: How do the concepts and relationships in Euclid's *Elements* relate to the world? Can one reconcile the infinite precision of

Euclid's propositions to the finite precision of the indirect

measurements that rely on those very geometric propositions?

How?

This dilemma can be summarized as an apparent paradox:

Geometry is about the world. But: 1) Geometric propositions have

infinite precision and 2) All of our measurements have finite

precision. So how can geometry be about the world?

Specifically, how should one look at concepts such as

straight line, circle, and triangle? Are they valid, meaningful

concepts? In the world we live in, is there really such a thing as a

straight line, a circle, or a triangle? How should one look at

measurements of lines, circles, and triangles? And what, if

anything, do Euclid's propositions and his arguments for those

propositions say about the world?

In addressing these questions, I will maintain that:

1.

Geometric shapes, such as lines, triangles, and circles, are

shapes that exist on earth.

2.

Euclid's propositions refer to shapes and relationships in

the world.
3.

Euclid’s arguments are valid: they reflect and capture

relationships that exist in the world.
These are three separate, but related, respects in which

geometry pertains to the world. Each requires separate discussion.

The last, especially, is the most mystifying, yet the most important.

And, on this point, to understand Euclid’s implicit method is to

understand his arguments, to understand how those arguments

relate to the world. For despite Euclid’s bad beginning, his work is

fundamentally sound.
This chapter is not a polemic against prevailing views. If

mathematics *means* anything, in any serious sense, it refers to aspects of the world and to relationships that exist in the world. So

the task of this chapter is simply to provide an account of, to

identify, the relationship of geometric knowledge to the world,

starting with its base in perception.
The overriding thesis of this chapter is that Euclidean

geometry is the rigorous study of shapes and spatial relationships

that exist on earth.

Timeline of Greek Mathematicians and

Philosophers

Euclid’s

Euclid's

Elements were a culmination of a long Greek

tradition, a tradition that was both mathematical and philosophical. To place Euclid within that tradition, I offer a timeline, as Figure 1, relating a number of important Greek philosophers and mathematicians during the classical period:[12]

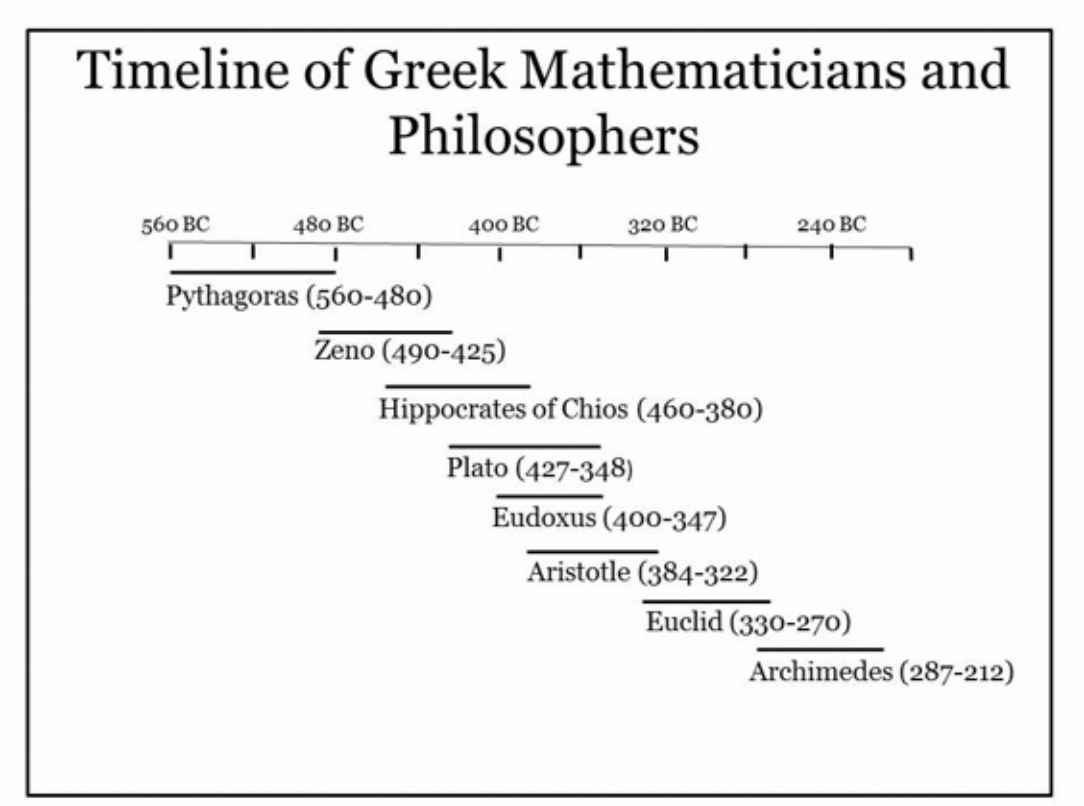


Figure 1

Though preceded by Thales, Pythagoras was the first major figure in Greek geometry. Pythagoras and his school are most noted for the celebrated and fundamental Pythagorean Theorem. But this is just one highlight of their reputed accomplishments. Most

significantly, the Pythagoreans are believed to have developed a

theory of proportion, entailing that the ratios of corresponding

sides of “similar triangles” (Triangles having the same angles, i.e., having the same shape) are equal.[13] The theory of proportion is fundamental. It provides the foundation of trigonometry upon

which we navigate the earth and measure the heavens. It is the cornerstone of indirect geometric measurement.[14]

But, unfortunately, the Pythagorean theory of proportion

made a critical assumption that turned out to be false. The theory

assumed that any two given lengths are commensurate, meaning

that both lengths can be expressed as whole multiples of some

common length. When the Pythagoreans said that all is number,

they did not have irrational numbers in mind: To the contrary,

when they said “number” they were speaking of positive integers.

So when the Pythagoreans discovered the existence of irrational

numbers, that the diagonal of a square, for example, is

incommensurate with its sides, their theory of proportion and even their broader world view became the casualties.[15]

As Charles Boyer observes, the natural way to have

salvaged the Pythagoreans’ theory of proportion would have been

through a limiting process, but that approach was shortly

discouraged by the need to answer the famous paradoxes of Zeno, all of them involving infinity.[16] For example, in perhaps the msot famous of his paradoxes,

Zeno argues that a faster runner (the

"hare") can never catch a slower one (the "tortoise"); that the hare, having given the tortoise a head start, can never catch up.[17] The hare can never catch up because, over and over again, every time

the hare reaches a point previously occupied by the tortoise, the

tortoise has moved on.

The Greeks believed that such arguments, no matter how

contrary to common sense, needed to be answered. But, in Zeno's

time, they had no answer.

Nonetheless, the Greek geometric tradition continued.

Hippocrates of Chios, credited with the first attempt to systematize

the "elements" of geometry, was, perhaps, the most notable

mathematician in the period following Pythagoras. The efforts by

Hippocrates to "square the circle" (i.e., to determine its area) led to

some promising related discoveries. Specifically, he was able to

square the lune, *i.e.* to measure the area of the shape between two intersecting circles curving in the same direction.[18]

Ultimately the classical Greek answer to the Pythagorean

dilemma of incommensurate ratios (in modern terms, irrational

numbers), as well as the Greek approach to limits (the "method of

exhaustion"), was provided by Eudoxus,[19] born after Plato but before the birth of Aristotle. Although none of Eudoxus's work

survives in written form, his work was essential to the theory of

proportion later presented in Euclid's books V and VI,[20] as I will discuss in later chapters.[21] It is a commonplace today that Eudoxus's theory of ratio anticipates Dedekind's approach to

irrational numbers, developed in the 19thcentury.[22] And Eudoxus's method of exhaustion, exploited by Euclid and, later, by

Archimedes[23] captures the key insight embodied in the modern theory of limits, precisely defined for the first time by Cauchy in the

early 19th century.

Euclid, then, stood at the

culmination of a

long

mathematical tradition. His *Elements* was the first truly successful systematization of Greek geometry. The *Elements* emerged in the wake of the flowering of Greek philosophy, epitomized by the work

of Plato and Aristotle. As one should expect, Euclid's work clearly

reflects the influence of both philosophers.

In that connection, to avoid any misunderstanding

regarding the Platonic influence, when I argue that Euclid's

Elements can and should be provided a non-Platonic interpretation and justification, I do not argue that Euclid would have agreed with

me nor do I deny the Platonic elements in his work. Rather I

maintain that he *should* have agreed with me, that the Platonic aspects of his *Elements* are not the essence of Euclid's method.

The final mathematician on this list is Archimedes,

properly regarded as the most original of the Greek mathematicians

properly regarded as the most original of the Greek mathematicians and, indeed, one of the greatest mathematicians of all time. Archimedes was the first to finally square the circle, and he went on from there to determine, with modern rigor, the volume and surface area of the sphere and the volume of a cone. Applying Eudoxus's method of exhaustion, Archimedes's methods in this regard anticipate the integral of Newton's calculus.[24]

Geometric Shapes

One does not grasp the concept of a "straight line" because someone has offered a definition. One knows what straight lines are because one has seen them, and has isolated them conceptually, say, by distinguishing them from crooked or curved lines. One has observed the straightness of straight lines and one has given them a name. In pointing to the relevant distinctions and similarities, one's definition of straightness is ostensive.

Nor does one chronologically begin with lines drawn on paper; one begins with shapes. One begins with circles, squares, and triangles; one begins with balls and cubes. A straight line is the edge of a rectangle (the shape of a book or a tabletop), the shape of a pencil, or the shape of a stretched string before it reappears as a line drawn on a sheet of paper. As one attends to shapes, one

observes that some shapes have straight edges. One recognizes straightness perceptually in such contexts. One perceives whether an edge or a line is changing direction; if it does, it isn't straight. If it doesn't change direction, it is straight.[25]

When one forms concepts of triangles, circles, and cubes, one is classifying actual shapes of actual objects within the environment, according to similarities and differences that one observes perceptually.

Consider, for example, a triangular shape. Triangles are first recognized perceptually. Typically, triangles in our environment are man-made. They include triangular drawing instruments, musical triangles, gables, pizza slices, structural elements in certain bridges and electrical towers, the arrangement of billiard balls at the beginning of a pool game, and the standard arrangement of bowling pins in a bowling game. Finally, a drawing of a triangle is, itself, a triangle.

Perception is the starting point and the base of our knowledge about geometric shapes and relationships. Euclid's *Elements* appeal to that base from the very beginning. His *Elements* show pictures of triangles, circles, and/or line segments on almost every page. Euclid appeals to our *perception* of his figures as an aid to grasping his arguments. And his propositions identify

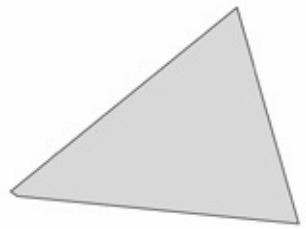
relationships among these figures. So the first thing to understand is the nature of these shapes. Let us continue to focus on triangles as a geometric shape that is both complex and relatively simple.

A triangle is a closed plane figure with three straight edges.

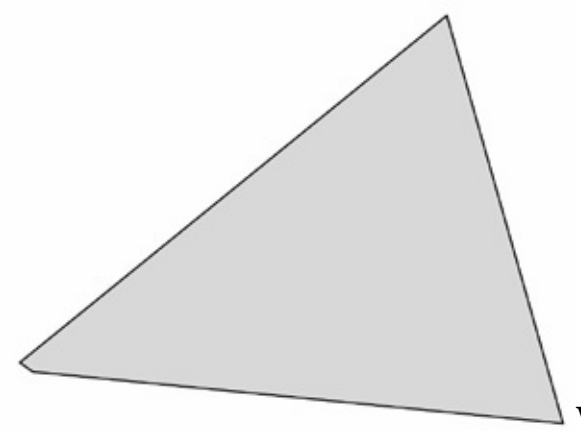
So when does something count as a triangle? Triangular objects have a multitude of characteristics. Many triangular objects, for example, have various colors. Some are red, some blue. But color is, obviously, not one's focus when one considers triangles: Not all aspects of an object are relevant to its shape. What makes something a triangle is the straightness and the number of its edges. Color doesn't matter; has nothing to do with whether or not a shape is triangular.

What about the other characteristics of triangular objects?

Let's take a harder example:

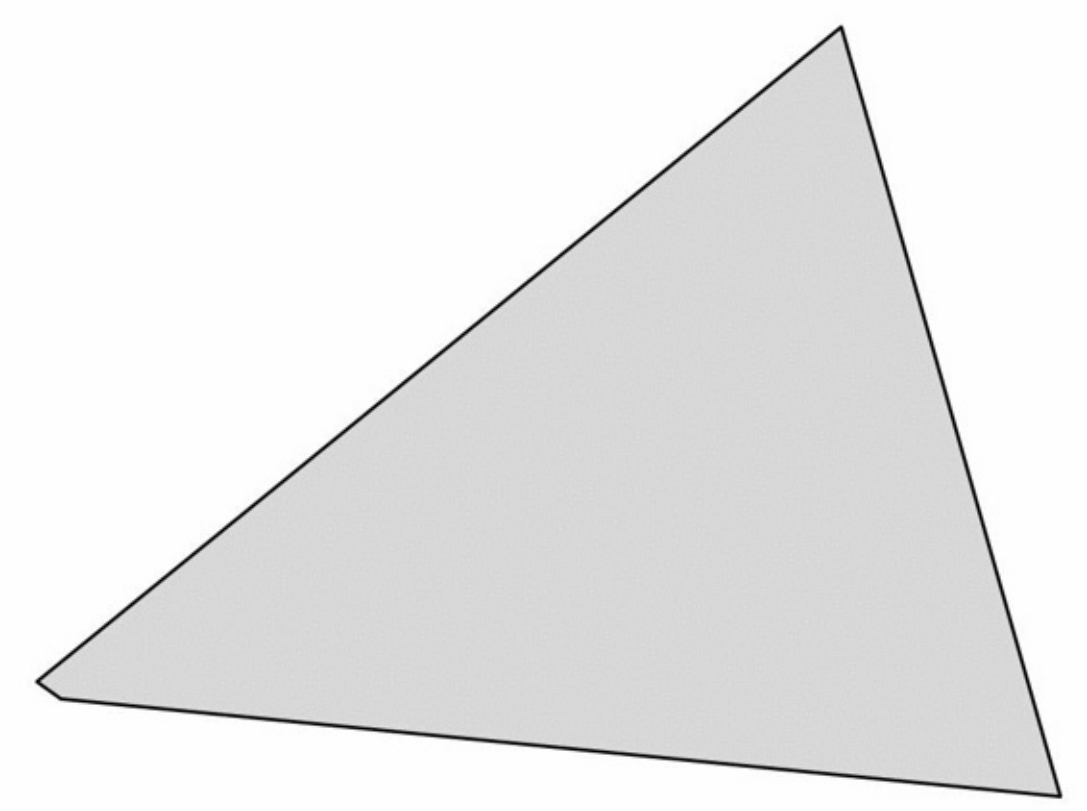
 Consider the following shape:

Is this a triangle?



What if I make it bigger?

Or still bigger?



At a certain point of magnification, for this particular

shape, one starts seeing a fourth side. So where does that leave us?

Arguably, each of these shapes has four sides. So, are none of them

triangles? Is, perhaps, the first example a triangle while the two enlargements are not? Is there some a priori point at which the fourth side is simply too big to ignore? It does us no good to say, "No they are all four-sided figures," if a further magnification will reveal a fifth side.

The question is far from academic. First, any triangle will exhibit imperfections if sufficiently magnified. The edges will be slightly curved or have slight knicks. The corners will be cut off, as in this example. Ultimately the triangles will resolve to distinct atoms, vibrating in some kind of stable equilibrium. If any imperfection, on whatever level of magnification, disqualifies something, per se, as a triangle, then there are no such things as triangles. There would be no such thing, indeed, as any particular geometric shape that we are able to name or identify.

Now we know that triangles exist. We formed the very concept from our perceptual observations of them. Our knowledge of more complex figures, including the analysis of four-sided figures, depends on our knowledge of triangles. Our ability even to frame the question, even to view a magnification of a shape as a *magnification* of that shape, depends on our knowledge of triangles. If there is no such thing as a triangle, our attempts to

analyze shape, even to criticize, cannot get off the ground.

So when is something a triangle? It depends on one's

standard of precision. In this example, if the fourth side is *relevant*, it's a quadrilateral (four sided figure). If the fourth side is

immaterial, it's a triangle. Whether something is a triangle depends upon one's standard of precision, of materiality. Since one's

standard depends, or should depend, upon one's specific cognitive

purpose, the same figure may be regarded as a triangle or as a

quadrilateral, depending on one's *standard* of precision.

One's cognitive purpose, one's standards of precision, one's

need to make certain distinctions, are all part of the context of one's

identifications. Where there is a continuum of possibilities, such

factors determine where one will draw the line, where one *should* draw the line. In this sense, the application of a concept, such as

"triangle", to concretes is contextual; it depends on one's context.[26]

For example, if one is interested in the number of calories

in a pizza slice, one would probably treat its shape as triangular,

despite fairly obvious differences between its shape and that, say, of

a triangular drawing instrument. But, conversely, a drawing

instrument with a broken corner would, as far that corner is

concerned, be unsuitable for its intended use as a triangular

drawing instrument.

So standard of precision is a question of relevance: What

So standard of precision is a question of relevance. What distinctions do you need to make? An immaterial imperfection is one that you don't have a reason to care about. A material imperfection is one that you *do* have a reason to care about. In either case, regardless of which judgment one makes, one is identifying an objective fact, either, say, that the fourth side is relevant or that it is not. If a fourth side isn't material, it is an objective fact that the figure is a triangle within one's standard of precision. If a fourth side is material, it is an equally objective fact that the figure is a quadrilateral, a four sided figure. The difference between the two cases is not in the shape of the object; that shape is assumed to be the same in either case. Rather, the difference is in the context, in the standard of precision appropriate to that context. In either case, the standard of precision is finite. If one regards the shape as quadrilateral rather than a triangle, one is just applying a finer standard than the one that saw it as a triangle. There remain further microscopic features that remain irrelevant to the more demanding context.

In sum, "A triangle has three straight sides" means:

1. Considered as a shape, there are three relevant sides
2. [illegible]

There is no relevant bending of any of the sides
3.

There are no relevant discontinuities in any of the sides

When Euclid *treats* the edges of triangles as if they were

perfectly straight and as if they were without width, his

mathematical *treatment* should not be taken to have metaphysical implications about the nature of physical triangles, nor should the

objects of his inquiry be taken as living in some separate

mathematical universe. Geometry is a specialized study that focuses

on certain attributes of objects considered in isolation from the

rest, not the study of idealized objects. “Triangle” is not about

disembodied shapes in some mathematical universe, lacking color

and substance, no more than “red” is about disembodied, shapeless

hues occupying a separate color universe. Rather, “triangle” is an

abstraction that zeros in on certain attributes of actual shapes of

actual objects and omits consideration of the other attributes of

those objects.[27] One’s application of “triangle” does not deny the reality of those other attributes. But one’s study of triangles can

ignore them precisely *because one’s knowledge about triangles*

applies to all triangular objects, regardless of the specifications of those other attributes.[28] All conceptual knowledge, including knowledge about triangles, is knowledge of the world.

Unless one looks at triangles in this way, one’s analysis of

geometric shapes cannot get off the ground. One analyzes the

geometric shapes cannot get off the ground. One analyzes the complex in relation to the simpler. Simple shapes such as triangles are the base for studying more complex figures. For example, one derives the area of a five-sided figure by applying the formula for the area of a triangle. One applies one's understanding of triangles to study more complex figures insofar as they differ materially from triangles.

But what determines relevance, materiality, or the appropriate standard of precision?

Suppose you are 10 seconds late to a meeting. Are you really late? Are start-times for meetings, under normal circumstances, even *specified* with that kind of precision? If not, what would it even mean to be 10 seconds late? Consider that the purpose of a meeting is to spend some time pursuing an agenda. To claim that ten seconds, according to the watch of one of the participants, were enough to make someone late would drop the context of the purpose and specificity of the rendezvous.[29] On the other hand suppose you are 10 seconds late swinging a bat. If your purpose is to hit an approaching ball, you are unbelievably late. A delay of one second would be too late! To say, "Well, I was only 10 seconds too late!" would *really* be dropping context! The appropriate standard of precision depends on the context. What is your purpose? What degree of precision is required to achieve that purpose? Classification of a figure as a

triangle or as something more complicated depends on one's *context* and on a *standard of precision* appropriate to that context.

Now consider the idea of a *perfect* triangle. Any triangle has microscopic imperfections. But if these imperfections are invisible, one normally says that the triangle is a perfect triangle.[30] But the key point is that *any* judgment of perfection involves and requires a *standard* of perfection. Normally visibility sets that standard, sets the required level of precision, but not always.

Sometimes more is needed. For example, a circular piston requires a degree of precision greater than the eye can distinguish. The geometry of semiconductors requires still greater precision and we are approaching the limits of what is actually possible. But, whatever the specific context, there will always be a limit to the precision that is actually available and the specific precision requirement in any particular case will always be finite. To ask for more is to attempt the impossible.

Measuring Triangles

One recognizes triangles perceptually as plane figures having three straight edges. One distinguishes them from other shapes with crooked or curved edges and from still other shapes that have a different number of straight edges. But, insofar as one

focuses on the shape, one does not distinguish them as to color. A red triangle is exactly as much a triangle as a blue triangle.

Nor does one distinguish triangles as to material. A wooden triangle is simply a triangle. Insofar as one views it as triangular, it is no different from a plastic triangle of the same size and shape. Insofar as one views them as triangles, the two shapes are identical.

Any entity has a multitude of attributes that all coexist with each other and are physically inseparable from each other. A triangular object must have some color and some material composition, but it may have any. The shape, the material composition, and the color are all present in every triangular object. But when one studies triangles, one focuses on the shape and on the measurable characteristics of that shape. And everything one learns about that kind of shape applies equally to all triangles, regardless of whatever color and whatever material composition any one of them might have.

If I say that triangles do not have color, I do not mean that triangular *objects* don't have color. I recognize that, generally speaking, triangular objects do have color. But the specific color doesn't matter, does not affect one's study of triangles. Color is one attribute; shape is another. The length of the edges is a measurable

characteristic of the *shape*; the color is not.

When one

studies triangles, one studies its *measurable characteristics as triangles*. One ignores color and focuses on the

lengths of the sides and the degree of the angles. When one attends

to color, one attends to the measurable characteristics of color,

namely hue, saturation, and intensity. One ignores the shape of

colored objects. The relationships among triangular shapes do not

depend on color; the relationships among colors do not depend on

shape.

But what about the thickness of its edges, the microscopic

or even visible crookedness of its edges, and the microscopic

discontinuities of the edges? These are measurements of *shape*. But are they measurements of *triangles*?

Now, in point of fact, when we compare

triangles, we

compare the lengths of their sides and the degrees of their angles.

We *ignore* their microscopic or irrelevant imperfections. We *ignore* the thickness, slight crookedness, and microscopic discontinuities

of the edges. Why?

These imperfections, were they relevant, would disqualify

the shape as triangular. Insofar as such characteristics relate to the

measurement of a triangle, what they actually measure is the extent

measurement of a triangle, what they actually measure is the extent

to which the

shape is

not a triangle. They are certainly

measurements of the shape. But they are only relevant to figures

that do not qualify as triangles.

We

qualify shapes as triangles insofar as imperfections,

such as the thickness of the edges, don't matter. And if something

doesn't matter then it really doesn't matter. If it doesn't matter then

one doesn't measure it. The *degree* of the imperfection is certainly relevant to *qualifying* a shape as a triangle. But qualifying a shape as a triangle *consists* in finding that its specific imperfections are irrelevant. Accordingly, when one *compares* two triangular shapes, one has already recognized that any imperfections in either of the

two triangles are *irrelevant*. One cannot, without contradiction or context-dropping, treat an *irrelevant* feature as if it were *relevant*.

In sum, when one studies triangles, one ignores irrelevant

imperfections just as one ignores color.

So mathematics studies the properties of triangles insofar

as they are triangles, insofar as they qualify as triangles within the

applicable precision requirements. Precision is not a feature of

triangles, as such. The Euclidean study of *triangles* does not, *per se*, include a study of precision. On the contrary, Euclid's *Elements* studies triangles insofar

as precision is not an issue. In the same sense that Euclidean triangles do not have color and that their edges do not have width, Euclidean triangles do not have precision. Within a particular context, precision requirements determine only whether a particular shape counts as a triangle. But the particular level of precision does not constitute an additional element of a triangle, like the lengths of the edges, that distinguishes and measures aspects of the triangle attribute.

In Ayn Rand's terms when one measures triangles, color is an omitted measurement. A triangular object must have some color, but it may have any. The particular color doesn't matter; it does not affect one's study of shape. In the same way, when one measures triangles, any microscopic or *irrelevant* imperfections of the triangle are omitted measurements. If these imperfections were *relevant*, we couldn't count them as triangles. But if something doesn't matter, one doesn't measure it. One omits it from one's analysis.[31]

The operation of Ayn Rand's principle of measurement omission is ubiquitous in mathematics and I will point it out repeatedly in this book. Ayn Rand introduces her concept of measurement omission when she observes:

"In order to form the concept "length," the child's

mind retains the attribute and omits its particular

measurements."[32]

This omission is a matter of implicit

method. Omission

does not imply non-existence. An abstraction integrates one's

awareness of a class of existents, including all of the characteristics of these existents.[33] To omit a consideration of certain specifications of its units is to apply a particular focus to these

units; it is not to expel these specifications from the abstraction, to

assign them to oblivion. As Ayn Rand puts it:

"Bear firmly in mind that the term "Measurements

omitted" does not mean, in this context, that

measurements are regarded as non-existent; it

means that

measurements exist, but are not

specified. That measurements *must* exist is an

essential part of the process. The principle is: the

relevant measurements must exist in some quantity, but may exist in any quantity."[34]

Concepts apply to real existents, to existents that have

countless differences within an essential similarity. To subsume

these existents under a single concept is to recognize and to focus

on characteristics that do not depend on these differences.

To return to my discussion of precision: Euclid's

Elements does not study precision, but there is such a study. One can

quantify *precision* as a separate study. For example, one can measure deviations from straightness; curvature is one such

measure. One can compare shapes that *don't* count as triangles to other shapes that *do* count as triangles and quantify the various respects in which they differ. But this *is* a separate study; a shape that is close to being a triangle, but is materially different from a

triangle, is not a triangle. Its comparison to a triangle is not a

comparison of two triangles; it is a comparison of two shapes. And

such a study, even a study of precision, is subject to its own

precision limits and builds upon the study of triangles. Indeed, the

comparison on a non-triangle to a triangle presupposes a standard

of precision in which one of the shapes being compared counts as a

triangle and the other one does not.

It may be helpful to contrast my view with that of John

Stuart Mill. Mill presents his perspective in the chapter

"Demonstration and Necessary Truths" of his *A System of Logic*.

Mill observes, for example, that,

"There exist no points without magnitude; no lines

without breadth, nor perfectly straight; no circles

with all their radii exactly equal, nor squares with

all their angles perfectly right."[35]

Mill deals with these observations by placing geometric objects in our imagination. He appeals to "one of the characteristic properties of geometric form–their capacity of being painted in the imagination with a distinctness equal to reality; in other words, the exact resemblance of our *ideas* of form to the *sensations that suggest them*."[36] [emphasis mine] Concerning actual objects, he says "… we feign them to be divested of all properties, except those which are material to our purpose, and in regard to which we design to consider them."[37] He ends by saying "… the conditions which qualify a real object to be the representative of its class are completely fulfilled by an object existing only in our fancy."[38]

Now superficially Mill's view might be taken to resemble my own. Mill, in effect, acknowledges that precision is finite and he also identifies the issue of materiality. But that is where the resemblance stops and he has it exactly backwards.

First, what is Mill's geometric object? Not something in the world viewed from a certain perspective, but a creature of the imagination that has been *suggested* by objects in the world.

Secondly, when Mill takes these objects in the world to "exactly resemble" the geometric objects of our imagination, he is comparing an existent with an idea. But a concept is not an object

to which one can compare one of its units.

Most fundamentally, Mill completely misses the referential character of concepts. The concept of a triangle does not refer to something in our imagination to which we can compare external objects; it refers to actual shapes in the world as viewed from a particular perspective, a perspective from which one attends to and measures certain characteristics isolated by the concept.

In sum,

Mill's view is characteristic of the representationalism of the British empiricists and it is a secular form of Platonism. It's like saying that horses in the world can be viewed as horses because they resemble your imaginary idea of a horse. Mill simply replaces Plato's world of Forms or Ideas by a world of the imagination.

To summarize this section, geometry studies triangles and other shapes. It studies and pertains to actual triangles on earth, not idealizations of triangles, limits of triangles, or imaginary triangles. Its concepts pertain specifically to those shapes.

Geometry is a specialized study: One focuses on what is relevant within a context and treats the other characteristics of its objects as omitted measurements.[39] The rest of this chapter will explore what geometric propositions say about those shapes and how the proofs

of those propositions apply to those shapes.

Mathematics as the Science of Measurement

I agree with Ayn Rand's characterization: "Mathematics is the science of measurement,"[40] and I believe that her characterization gets at the heart of both the nature and the purpose of mathematics.

But this has never been the standard view. Indeed, it will strike many as extremely odd, as trivializing an enormously abstract and complex subject. Nonetheless, a major goal of this book is to show just how the abstract complexity of mathematics should be understood from the measurement perspective.

As a point of reference, it is important to see Ayn Rand's characterization of mathematics from a historical perspective. In the Classical period, as I have already indicated, Plato held that mathematics was about mathematical entities living in an abstract realm, the world of Ideas.[41] Aristotle, coming from a more worldly perspective, held that mathematics was the science of quantity.[42] Aristotle's view survived into the modern period but began to be challenged by both mathematical and philosophical developments during the nineteenth century. As the mathematical developments seemed to progressively transcend the traditional

understanding of quantity, Aristotle's view was ultimately abandoned.[43] In the words of Moritz Epple, the 19th century witnessed "the end of the paradigm of the science of quantity."[44]

Aristotle's view is the closest to Ayn Rand's in any historical period. But there is an important difference between the two views. Namely, Ayn Rand's perspective is epistemological: Measurement involves a process of establishing a relationship, a relationship of the attribute one is measuring to a standard. Measurement involves mental connections, connections based on quantitative *relationships* observed in the world. Mathematics is not idle contemplation; it has a cognitive function: integrating the conceptual and perceptual realms, establishing quantitative relationships to simultaneously differentiate and connect the entire spectrum of knowledge.

In contrast, Aristotle's perspective is metaphysical:

> Aristotle's quantity exists in the world as a fundamental category. It is the quantities themselves, not their relationships to other quantities that require special study. And quantity, for, Aristotle and his contemporaries referred either to pluralities or to that which was divisible.[45]

Mathematicians during the 19th and 20th centuries expanded the mathematical field in ways that the Greek civilization could not have imagined. These developments were all driven by

the needs of measurement, of solving equations and identifying geometric relationships. As a pursuit, mathematicians continued as they always had, devising new techniques and developing new concepts to solve each new problem as it rose from the ashes of previous problems already laid to rest.

Yet their investigations led in unexpected directions: to nonEuclidean geometry, to algebraic structures such as groups and rings, to entirely new disciplines such as algebraic topology. Nonetheless, one could have adapted the measurement paradigm to embrace and provide a perspective, indeed, illumination, on these developments. And it would have been possible, though harder and less natural to similarly adapt the idea of quantity.

These developments were driven by methodological considerations, ultimately by the requirements of indirect measurement, and are profitably understood as such.

But German mathematicians, in an environment shaped especially by Kantian ideas, chose a different path, one the rest of the world soon followed. In this environment Aristotle's conception, mathematics as the study of quantity was easily abandoned. Rand's conception of quantitative relationships would

abandoned. Rand's conception of quantitative relationships would have been easier to adapt, harder to abandon, but, in any event, was not available as an alternative.

In my view, the study of quantity should be pursued as a study of quantitative *relationships*, relationships among similar characteristics. And this is the key ingredient of measurement.

Aristotle's conception tends to divert one from this point; Rand's definition of mathematics embraces it.

The new conceptions of mathematics that arose during the nineteenth and early twentieth centuries included David Hilbert's view of mathematics as the study of formal systems,[46] Frege's and Bertrand Russell's attempt to reduce mathematics to symbolic logic,[47] and the development of set theory as a purported foundation of mathematics.[48] This latter set theoretic approach has many fathers, but a key milestone on that path was Georg Cantor's conception of an actual, completed infinity.[49] In total, these developments go far beyond the abandonment of Aristotle's quantity. They all are, to one degree or another, an abandonment of the view that mathematics is referential, that mathematics is about the world and pertains to the world. For further elaboration and, specifically, for the contrast between my own view of sets and the modern perspective, see Chapter 6.

As already noted, realism in mathematics is taken today, among philosophers of mathematics, as almost a synonym for Platonism.[50] That mathematics might refer to aspects of the world is generally dismissed, primarily because of issues involving

infinity, specifically in light of the ubiquity of infinity in mathematics.[51]

Nonetheless there have been a number of modern attempts to define and defend various versions of mathematical realism. I applaud such efforts, though I cannot offer a survey, much less a critique, of this work here. I can only indicate a broad contrast to my own approach and I restrict myself to one illustration. Namely, consider Michael Resnik's *Mathematics as a Science of Patterns.* As he puts his thesis, "The ontological component of my realism is a form of structuralism. Mathematical objects are featureless, abstract positions in structures (or more suggestively, patterns); my paradigm mathematical objects are geometric points, whose identities are fixed only through their relationships to each other."[52]

Much later, he elaborates:

"The objects of mathematics, that is, the entities which our mathematical constants and quantifiers denote, are themselves atoms, structureless points, or positions in structures. And as such they have no identity or distinguishing features outside a structure."[53]

In stating this thesis, Resnik is affirming that, in some abstract sense, mathematics has an object. But, at least in the sense

I intend, he is not saying that mathematics pertains to the world.

In my view, Ayn Rand's characterization of mathematics gets to the heart of the subject. I maintain, in this book, that to understand how mathematics relates to the world, one must understand how mathematics relates to measurement. Indeed, the key to a full understanding and appreciation of mathematics, from both a philosophical and mathematical perspective and on all levels of mathematical abstraction, is to understand its concepts, propositions, and demonstrations in relation to measurement. In this chapter I apply that perspective to Euclid's *Elements*.

What is Measurement?

In Ayn Rand's definition,

"Measurement is the identification of a relationship–a quantitative relationship established by means of a standard that serves as a unit."[54]

The purpose of measurement is to extend and objectify our grasp of the world *beyond* what we can directly perceive[55] by identifying quantitative relationships to what we *can* directly perceive. One of those means is the application of a universal standard, a universal reference point to which all quantities of a particular kind relate.[56]

Establishing quantitative relationships, however, is something more general than measurement in Rand's sense and establishing a quantitative relationship between two quantities is possible without reference to a standard. For example, "This pencil is longer than that pencil," identifies a quantitative relationship but makes no reference to a standard and is not, in this sense, a measurement according to Ayn Rand's definition. On the other hand, a determination that, "This board is 5 feet long," counts as a measurement because it relates the length of the board to a standard, namely to a foot.

So what does geometry measure? It measures shapes and spatial relationships. In particular, geometry measures triangles, distances, directions, areas, and volumes.

Measurement is complex. The measurement identification entailed in Ayn Rand's definition represents a culmination of a process. But measurement, in the full sense, presupposes that:

1. Standards have already been chosen.
2. Subdivisions are available and there is a process in place for making further subdivisions.
3.

There is an inventory of direct and indirect methods to

relate an object of measurement to a standard.
4.

There is a general way to express the results of a

measurement.

Each of these elements requires a separate discovery or,

indeed, a series of discoveries. Ayn Rand's

definition

of

measurement pertains to the finished product that embodies all

these separate discoveries. Euclid's *Elements*, in particular, is a systematic study and integration of such discoveries, of geometric

relationships that can help relate an object of measurement to a

standard.

This chapter studies the roots of geometric measurement in

Euclid.

Indirect Measurement

Measurement begins by making quantitative comparisons,

relating differences in degree. For example one judges that this

pencil is longer than that pencil or that my right hand has the same

number of fingers as my left hand. These comparisons do not

require measuring either pencil or counting the fingers on either

require measuring either pencil or counting the fingers on either hand. But they do reveal quantitative relationships. Identifying quantitative relationships is the fundamental underpinning of measurement.

Geometry offers the ability to transcend the limits of direct comparison. It provides powerful tools for establishing quantitative relationships through indirect means. It provides an essential foundation for indirect measurement. By geometrically-inspired calculations, Eratosthenes, in 200 BC, measured the angle of the sun's rays at noon to find the circumference of the earth.[57] And we use geometry today to measure the distances of stars in space and the arrangement of atoms in crystals.

In my usage, an indirect measurement is one that is not, itself, a direct measurement, but derives from more direct measurements or relies on measuring and calibrating a causal consequence of the attribute being measured.[58]

Simple addition is an elementary case of indirect measurement. If one has 12 dimes in one pocket and 14 in the other, one does not need to count the aggregate to know that one has a total of 26 dimes. One uses the laws of addition, i.e., mathematical relationships, to quantify the aggregate. Because of elementary arithmetic, a millionaire knows he is a millionaire

without ever counting to a million.

Arithmetic starts as a means of indirect measurement, but why do we need indirect measurement so pervasively? What is the general pattern?

The simplest measurements involve direct perception. One counts to determine multiplicity, uses tape measures to measure length and distance, plumb bobs to determine the vertical, carpenters' levels to determine the horizontal, and protractors to measure angles.

But consider the measurement of weight by an electronic scale, of the speed of one's car by a speedometer, of elapsed time by a stop watch, or of the voltage difference in a battery by a voltmeter. These are the sorts of measurement tools one needs once one goes beyond the simple comparisons available to perception.

Measurements involving gauges, such as these, typically follow the following pattern: A causal sequence connects the attribute being measured with a reading on a dial, a reading that reflects a calibration of the effect of the attribute on the measuring device. The causal sequence, specifiable mathematically, connects the attribute being measured with the reading on the dial: Behind the machine stands physics and mathematics. A measurement made by

a machine is an indirect measurement.

This chapter will examine the

mathematical underpinning

of indirect measurement and exhibit indirect measurement as the

key to understanding both the content of Euclid's propositions and

the method of his arguments. For every proposition in Euclid's

Elements states something about indirect measurement. Every proof appeals to a series of measurements, measurements that,

taken as a whole, suffice to indirectly establish the relationship

asserted in the proposition.

As a useful example of indirect measurement, consider a

flagpole depicted in Figure 2. Suppose that, at some particular time

of day, the flagpole casts a shadow ten feet long. Suppose, further,

that, at exactly the same time, a nearby six-foot man casts a shadow

three feet long. How high is the flagpole?

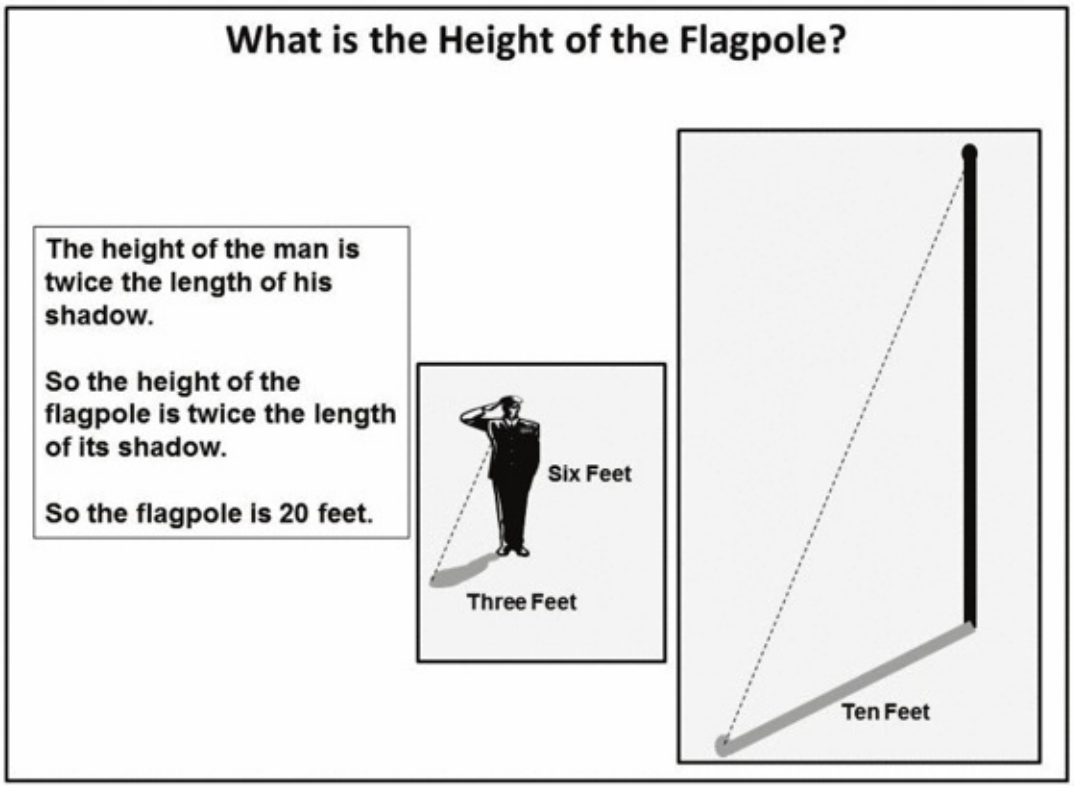


Figure 2

Now one may know that the length of the shadow is proportional to the height. One may, perhaps, reason: “Triple the height; triple the shadow.” So, one observes, the height of the man is twice the length of his shadow. So the height of the flagpole must be twice the length of its shadow. The flagpole’s shadow is ten feet. So the flagpole must be 20 feet.

But how does one know that the length of the shadow is proportional to the height? It may seem obvious, but the discovery and validation of this principle was a high point of Greek mathematics, one with a difficult history. The ancient Greek school

of Pythagoras made an early, not fully successful, effort to establish the relationship in approximately 500 BC.[59] A century later, Eudoxus found the key to a fully successful demonstration. Euclid

presents a rigorous account, based on the work of Eudoxus, in Books V and VI of the *Elements*.[60]

Only Euclid didn't quite put it that way. He said, rather,

that the ratios of corresponding sides of similar triangles are

equal.[61] Now to say that triangles are similar is to say that they have the same shape, but that one may be bigger than the other.

Triangles have the same shape when they can be matched up so

that corresponding angles are equal.[62]

So suppose two triangles have the same shape. According

to Euclid's proposition, if a side of one triangle is twice the length of

its base, then the corresponding side of the other triangle must be

twice the base of *its* triangle.

Ok, where is the triangle in my example of the flagpole?

The answer is depicted in Figure 2. The man and his shadow form

two sides of a triangle. The third side is supplied by the rays of the

sun: A sunray just missing the top of the man's head will hit a point

just beyond the shadow. To put it another way: A line drawn from

the tip of the shadow to the top of the man's head points directly to

the sun. That line is the third side of the triangle. The angle at the

bottom represents the angle with which the sun's rays are striking

the ground.

In the same way, the flag pole and its shadow are also two

sides of a triangle. Corresponding angles are equal; the sun's rays

are all parallel and both the man and the flagpole are presumed to

be standing straight. So the triangles have the same shape.

Therefore, the shadow of the flagpole is to the shadow of the man as

the height of the flagpole is to the height of the man. From which it

follows that the length of the shadow is proportionate to the height

of the shadow caster.

I said earlier, in regard to gauges, that a "causal sequence,

specifiable mathematically, connects the attribute being measured

with the reading on the dial." These causal elements are observable

in the flagpole example. First, the *brightness* of the sunlit region surrounding the flagpole shadow is caused by the sun's rays. Or,

alternatively, the *shadow* is caused by the flagpole obstructing the rays of the sun. The *contrast* between the two regions, the pointer on the dial upon which we rely for our measurement, depends

jointly on the nature of the sun and the nature of the flagpole.

Specifically, the sun's rays propagate in straight lines. And those

rays that fall on the earth propagate in *parallel* straight lines, due to the great distance of the sun from the earth. The flagpole, for its

part, blocks the sun's rays and the shape of the shadow is

determined by the shape of the flagpole.

determined by the shape of the flagpole.

As a result of all of these causal elements, the length of the shadow is a manifestation of the height of the flagpole. The shadow provides evidence of the height of the flagpole: the taller the flagpole, the longer the shadow. The shadow acts as a gauge of the attribute being measured: the height of the flagpole.

Yet, so far, it is a gauge without a scale. The measuring instrument is the causal sequence connecting the sun, the flagpole, and the ground upon which the shadow falls. But that measuring instrument has not yet been calibrated.

For that, one needs to compare the reading on the ground with an appropriate standard. One needs a reference shadow for which the corresponding height is already known. That function is served by the six-foot man with the three-foot shadow.

But even that is not quite enough: One still needs to know how the length of the shadow varies with the height. Not all causal relationships, after all, are easy to quantify. But, in this case, geometry provides an answer that has already been noted: The length of the shadow is proportionate to the height of the shadow caster.

In summary, this calculation relies on such physical factors as, first, that the sun's rays propagate in straight lines, second, that

as, first, that the sun's rays propagate in straight lines, second, that the sun's rays are parallel, and, third, that both the flagpole and the man block the sun's rays. It also relies on the fact that one has already measured the height of the man and the lengths of the shadows. But the final step in this measurement provides the mathematical foundation that makes the rest relevant. That step is the application of the geometric proposition that when triangles are similar (have the same angles), the ratios of corresponding sides are equal. Taken as a whole, this process is an indirect measurement of the height of the flagpole.

As this example illustrates, mathematical relationships such as geometric relationships and quantitative expressions of causal relationships, are an essential underpinning of indirect measurement.

The process I have described embodies key features of indirect measurement taken generally. Like all measurement, it also offers finite precision: 20 feet plus or minus. What are some of the factors that may limit its precision?

There are many. It's possible, for example that the ground is not entirely level or that the flagpole isn't completely straight or completely vertical. The flagpole may be elevated above the general

level of the surrounding ground. Measuring the length of the shadow is subject to various precision limits, including an estimation of the point inside the base of the flagpole from which the measurement should start and a second estimation of which point on the shadow's extremity will provide the most accurate measurement.

There are also calibration issues. A six foot man is not the ideal shadow caster for this purpose; a vertical six-foot stick would be better and the narrower the stick, within limits, the better the determination. All of the issues with measuring the flagpole's shadow apply, as well, to measuring the man's shadow. Finally, one last measurement is needed, subject to its own precision limits: the man's height.

But there are no mathematical limitations. Short of human error, there is no additional error introduced by the mathematical calculation, which can be carried out to any required precision.

In simplest essence, this measurement of the height of the flagpole requires three direct measurements, specifically of the man's height and of the two shadows. And it requires one mathematical calculation. All three direct measurements are subject to specific precision limits. But whatever those limits may

be, the mathematics can derive the most accurate final measurement that is available under the circumstances. And, to the extent that the respective *precision* limits of the component measurements have been identified, the mathematics can calculate the precision of the final result.

The role of geometry in the indirect measurement of physical quantities is not always as transparent as it was in the flagpole example. However, physical measurements necessarily involve objects that are related spatially and have a geometric structure. These geometric relationships are part of the theater in which all physical phenomena takes place and our knowledge of geometry is embedded in the mathematics that we apply to the phenomena. For example, the operation of a balance scale reflects the way that physical forces interact, but it also depends on the geometric relationship between the lever arm and its fulcrum.

My points about precision apply generally, as well. For a balance scale, the lever arm, on a microscopic level, won't be completely straight. The fulcrum has a finite extension. And so on. But the mathematics, as in the flag pole example, will provide, and can quantify, whatever level of precision is needed within the physical constraints of a particular context.

Context and Precision

As I have discussed, precision is finite and the appropriate standard of precision is contextual. And a key aspect of that context is an answer to the question, “Why do you care?” For example, suppose your intent is to paint the flagpole with a very expensive paint. You want to make sure you have enough paint, but you don’t want to have too much left over. You may be looking for an estimate of the height within, say, six inches. Yet your measurement based on the length of its shadow may only be accurate within a couple of feet, significantly outside your precision requirement. But now suppose you have a 20 foot ladder and a 25 foot tape measure. In such a case, the precision of your initial estimate is more than adequate to the task because you have now established that your ladder and tape measure are all that you need to *refine* your initial measurement to meet your precision requirement. So your initial measurement, serving as an initial estimate, is perfectly precise. It fully meets the precision requirement of an *initial estimate* by establishing the feasibility of the *subsequent refinement*.

In general, if one needs to refine a measurement, one may refine the measuring instrument, which would be difficult in this particular example. Or one finds a more accurate way of measuring.

If a principle applies to a concrete, as in this case, then that

concrete is one of the units of the principle, part of the meaning of the principle. There is an exact parallel here to Ayn Rand's observation that the meaning of a concept consists of its units. In the same way, the meaning of a principle consists of its units, of the actual and potential concretes to which it applies.[63]

In general, mathematics cannot guarantee that a desired level of precision is achievable. But it can guarantee that if the *physical requirements* are met, the desired precision will be achieved. In this precise sense, the calculation is universally valid.

The principle is that if X, Y, Z, *etc.* are all precise enough, then the application of the mathematical principle will satisfy the precision requirements of the particular measurement.

Notice the parallelism between my discussion of triangles and measurement. In the case of triangles and other geometric shapes, context determines relevance. Context determines what counts as an imperfection, which imperfections are relevant. In the case of measurement, context determines the precision requirement, how close your measurement needs to be in order to constitute a precise measurement. The job of mathematics is not to make infinite precision possible. It is to accommodate any *particular* precision requirement that might be needed, some day, for some

reason, in any particular instance. A mathematical

principle that meets this demand applies equally to all qualifying

instances.

Geometric Propositions: Meaning and Precision

Euclid's propositions are about the world. In the flagpole

example, I applied the proposition: *Ratios of corresponding sides of similar triangles are equal*. Similar triangles are triangles for

which corresponding angles are equal. But what does it *mean* for two angles to be equal?

On the basis of everything I have said, there is only one

thing this could possibly mean: Two quantities are equal if there is

no material difference between them. So Euclid's proposition says that if there is no material difference between corresponding angles

then there is also no material difference between the various ratios

of corresponding sides. And recall, once again, that a material

difference is a difference that one has a reason to care about.

With this understanding, the proposition applies, as in the

flagpole example, to all pairs of similar triangles that, within their

particular context, satisfy the criteria of the proposition within the

applicable precision requirement. For example, suppose that the

flagpole height needs to be known within six inches. Addressing

flagpole height needs to be known within six inches. Addressing this need imposes definite requirements upon how accurately one measures the lengths of the shadows and the height of the man. It imposes further requirements on how similar the similar triangles need to be. But if those requirements are met, the final result will have the required accuracy.

The proposition applies to all contexts for which the physical requirements of the required level of precision have been met. In *any* such context, the proposition means that the ratios *are* equal, equal within the required level of precision. If one's final measurement requires precision within 5%, it might be necessary to measure the inputs to precision of 1%. But *some* degree of precision will suffice to meet the requirement.

The mathematics cannot guarantee that a desired level of precision will be physically achievable in a particular context. But it *can* guarantee that if the physical requirements are indeed met, the final result will be accurate within six inches, or half an inch, or, indeed, within any specific finite precision one might demand for the final result.

In sum, a geometric proposition applies equally to all situations of a particular type. It applies to an entire category of physical situations. In each case, the proposition provides an answer that will hold within the specific required precision

whenever the physical demands of that precision can be met.

To further explore the issue of materiality, introducing a

possible complication, consider a second example. A triangle is called isosceles when two of its sides are equal.[64] Euclid states that, for any isosceles triangle, the angles opposite the equal sides are

equal. Conversely, any triangle containing two equal angles is

isosceles.[65]

What does this statement about triangles actually say about

triangles, about actual physical triangles? Do isosceles triangles

even exist?

This second question is really no different than: Do

triangles exist? The answer is yes. An isosceles triangle is a triangle

for which, within a specified particular context, there is no relevant

difference between two of its sides.

The statement says that if there is no relevant difference

between the sides, then there is also no relevant difference between

its angles. One can also put it another way: To the extent it's a

triangle, any difference in the angles is attributable to a difference

in the sides. And a good way to look at either formulation is the way

I did with the flagpole example: No matter how close one requires

the two angles to be, one can guarantee it by making the sides

sufficiently straight and making the lengths of the two opposite

sides sufficiently close.

Now it might happen that a difference in the lengths of the

sides would be detectable while the difference in the angles would

not be detectable or vice versa. Lengths and angles, after all, are

different kinds of quantities and might be directly measurable with

different levels of precision. But in such cases, the difference, say, in

the lengths of the sides is *evidence* that the angles themselves are unequal even when this difference cannot be distinguished by a

more direct measurement of the angles. Armed with Euclid's

theorem, the inequality of the lengths of the sides provides an

indirect measurement of the relationship between the angles.

Now suppose that one can directly measure the angles, but

not the sides. In this case the proposition implies that, if the angles

are equal, this equality is evidence that the sides are equal. In such

a case, measuring the relationship of the angles *is* the most precise available measurement of the relationship between the edges.

Without such a warrant, the theorem would be useless. The

very point of relating angles and lengths is to provide a way to

measure quantities *indirectly* when one cannot measure them directly, by relying on mathematical relationships to quantities that

one *can* measure directly. When a sailor navigates by determining the directions from his location to the fixed stars, he is counting on

various mathematical relationships to establish his position,

mathematical relationships that relate distances and angles.

My examples illustrate how Euclid's propositions do, in fact, apply to the world, indeed, how his propositions have always been applied. In practice, there has never been a mystery in how one applies geometric knowledge and such applications preceded Euclid's brilliant integration.

I have so far maintained that mathematical concepts are derived from reality and refer to reality. I have discussed what geometric propositions actually mean about the world. But what about Euclid's *arguments* for his propositions?

If Euclid appears to be writing about ideal lines, circles and triangles, how can his arguments apply to real lines, circles and triangles? How can his arguments pertain to the world? Are his logical arguments really still valid if one applies those logical arguments to actual, real-world lines, circles, and triangles?

One cannot omit an answer to this question. My thesis demands an account of what the steps in Euclid's arguments actually mean. What do they say about and how do they apply the world? To this subject I turn.

How does Euclid Measure?

The rest of this chapter will pursue three basic questions

The rest of this chapter will pursue three basic questions concerning Euclid's *Elements*. Where does he start? What are his tools? How does he use them?

In outline, Euclid starts with "postulates" and "common notions". In my view, these provide the tools, the primitive measurements, to prove his propositions. I will first examine these postulates and common notions as they relate to measurement. And then study their use in proving the propositions.

Euclid's *Elements* embodies a highly stylized, abstract approach to measurement. Euclid does not use yard sticks or tape measures. He does not say things like: "Two sides of this triangle are each 5 inches therefore they are equal." Rather, he says things like, "Suppose that two sides of this triangle are equal. Then their opposite angles, whatever they might be, are equal, as well." Such a statement is an abstract formulation of a universal relationship applying to all triangles as such, independent of their specific measurements.

Every step of every proof of every proposition in Euclid is an abstract measurement or is part of one. Every construction is a process of measurement; every proposition expresses a mathematical relationship: a comparison of two or more distinct

quantities. Every comparison is either a direct measurement or it is an indirect measurement. But the indirect measurements already reflect a chain of direct comparisons and direct measurements. In a sense to be elaborated, Euclid's method is to present a recipe for a series of abstract measurements that establish a proposition. Measurement in Euclid consists in finding or specifying quantitative relationships.

Identifying Quantitative Relationships

It's important to understand the relationship of the wider concept, *identifying quantitative relationships*, to *measurement* (in the full sense of establishing the relationship to a unit).

Identifying quantitative relationships can be done without numbers or units It is a wider concept than measurement; it precedes measurement; it is presupposed, in general, in the act of finding a quantitative relationship to a *standard* (i.e., any act of *measurement*); it is very often a step in measurement; and, as used in Euclid, it is an abstract form of measurement.

First, the act of identifying quantitative relationships neither requires nor presupposes numbers nor units. Indeed, we make nonnumerical quantitative judgments routinely. To judge visually that Tom is taller than Mary, that a feather will feel lighter than a silver dollar, that one light is brighter than another, or to

point in the direction of a circling hawk are all to make a nonnumerical quantitative judgement.

Identifying quantitative

relationships without numbers or units is an everyday activity.

Although every numerical measurement establishes a

quantitative relationship, the reverse is not true, as the previous

examples indicate. “[I]dentification of a quantitative relationship”

is not the differentia of Ayn Rand’s definition of measurement; it is

the *genus* .

One identifies quantitative relationships before one starts

relating things to standards. One compares multitudes visually, or

by other means, before one learns to count. One compares lengths

before one compares the length of an object to a tape measure. And

any act of establishing a quantitative relationship to a standard

presupposes a

more general ability to establish quantitative

relationships.

The measurement of the flagpole by means of its shadow

relied on a general quantitative relationship relating similar

triangles. That universal relationship of similar triangles does not

depend on one’s choice of a standard of measurement and the

statement of the principle makes no mention of one. And this case

is typical of indirect measurements. Indirect measurements typically involve appeals to quantitative relationships that apply generally, across a wide spectrum of concrete cases. Such an appeal, whether it's identified explicitly or is only implicit, is a step, often a key step in the overall measurement process.

Finally, Euclid's identifications of quantitative relationships embody an abstract form of measurement. In any Euclidean argument, the specific standard units, numerical measurements, and context are irrelevant, are, in Ayn Rand's terms, "omitted measurements". For a Euclidean argument applies to an open-ended range of concretes, regardless of the particular standards, numerical measurements or specific context of each concrete that is included in that range of concretes. Euclidean propositions are all statements of quantitative relationships that transcend specific concrete measurements.

I have already discussed the issue of finite precision as it applies to the statements of Euclidean propositions, to the way one should think of relationships of equality and inequality, and to the way one should view the universality of these propositions. The same principles apply to measurement or to a series of

measurements. It applies to abstract measurement. And this is important to my broader argument because of my contention that every Euclidean argument, despite its deductive form, reduces to a series of connected abstract measurements.

To make that broader argument, the main remaining burden of this chapter will be to explain just how the steps in Euclid's *arguments* express quantitative relationships, how they constitute abstract measurements. But one word of warning regarding terminology: Although the best characterization of these steps is the term "abstract measurement," I will frequently, for simplicity of expression, use the term "measurement" to describe these steps. Whenever I do that, it should be taken to refer to the broader phenomenon of abstract measurement.

Straight-edge and Compass Constructions in Euclid

Neither straight edges nor compasses appear in the *Elements*. But when Euclid proposes drawing a line connecting two points, he appeals to the use of a straight edge. When Euclid proposes extending a line that has already been drawn, he appeals to the use of a straight edge. When Euclid proposes drawing a circle with a given radius around a point, he appeals to the use of a

compass.

Euclid systematically appeals, implicitly, to the use of straight edges and compasses, or at least of some kind of device to achieve the same ends. So to appreciate and understand Euclid's arguments, to understand their relationship to the world, one begins by identifying the essential nature of these constructions.

Euclid uses lines and circles in two related respects: first, to *construct* geometric figures to various specifications and, second, to identify geometric relationships by means that ultimately appeal to the nature of lines and circles. Lines and circles, and also angles, are Euclid's means of measurement.

Circles, Straight Lines, and Angles

Euclid constructs circles, straight lines, and, derivatively, angles. What exactly do these constructions measure?

First, what is a circle? A circle is characterized by the fact that every point on its circumference has the same distance from a central point. When Euclid posits that a circle can be drawn of any prescribed radius from any designated point, he refers to the possibility of measuring out a length of any prescribed amount, in any direction, from any particular point in the universe. Circles do not directly compare lengths of objects residing in different places. But circles do provide a direct way to compare a distance in one

But circles do provide a direct way to compare a distance in one

direction to a distance in another direction, as long as both

distances start from the same point.
A circle, in Euclid, therefore, functions as a measurement

of distance. As a measurement, it performs two essential tasks:

First, it provides a visual indication of distance from a central point.

Second, it provides a way of identifying a point in any required

direction having that distance from that central point.
Second, what is a straight line? A straight line is a line that

does not bend or curve, a line that continues in a single direction.[66] The use of an arrow to indicate a direction to something is a direct

expression of where something is, of where it is in relation to where

you are now. So a straight line, in Euclid, functions to specify

direction. As a measurement, a straight line performs two essential

tasks: First, it provides a perceptual identification of a direction

and, second, provides a way to find other points in that direction

from a point of origin.
When Euclid says that any straight line can be extended,[67] he is saying that one can continue moving in any direction that one

has specified or chosen. When he says that any two points can be connected by a unique straight line,[68] he is saying that there is a line of sight connecting any two points in the universe; that any

point B is related directionally to any other point A. He is saying

that one can go from anywhere to anywhere else by finding the

right direction and sticking to it.

right direction and sticking to it.
Although Euclid was clearly aware that he was using circles

to measure distance, it is less clear, and certainly not explicit, that

he was using straight lines to measure direction. My contention is

that, nonetheless, in the nature of the case, he was, in fact,

measuring direction and by means of straight lines.
But Euclid does not speak of direction except in the

following restricted sense: If a point divides a line segment, he

speaks of the two directions along the line from the particular

point.[69] In that case, Euclid's distinction is clear and directly perceptual. But he does not use the term, *direction*, more generally.

So, in light of the importance I place on Euclid's measurement of

direction, this caveat should be understood at the outset.
What is an angle? An angle consists in the difference in

direction of two intersecting lines at the point of intersection, as

manifested in the amount of turning required to rotate, at that

point, from one direction to the other direction. A small angle

represents a small amount of turning; a larger angle represents a

larger amount of turning. To continue turning in the same direction

is always to increase the total amount of turning from a smaller to a

larger degree. A rotation can be thought of as a continuous change

in direction.
When an amount of rotation is reflected in two line

segments that lay out the starting direction and ending direction of

segments that lay out the starting direction and ending direction of

the rotation, the two intersecting straight lines provide a concrete,

visual indication of the difference in the two directions. In this

sense, an angle measures a difference in direction at a point. When

one says that two angles are equal, one identifies that they embody

the same amount of turning, the same difference in direction.[70]
I am not yet speaking of numerical measurement of angles;

rather, I am speaking of angles considered as physical magnitudes.

However, it's important to notice that a *direction*, as such, is *not* a magnitude.[71]
To measure direction, one needs to *start* with a *particular* direction, with a standard direction that has been identified *perceptually*. In effect, to indicate a particular direction you ultimately have to point.
However, a *difference* of two directions in a plane *is* a magnitude. On the one hand, it does not make sense to say that one

direction is three times a different direction. But it does make sense

to say that one angle, one amount of rotation, is three times a

second angle. Once one standard direction from a point in a plane

has been perceptually identified, one can specify any other

direction (lying in the plane) from that point by an amount of

rotation from the standard direction. One is using a kind of

magnitude, an angle, to specify or measure a direction, by

specifying its relationship to a standard direction. However, the use

of a magnitude in the process of *measuring* direction does not make direction, as such, a magnitude.
In sum, Euclid appeals to our ability to *mark* a distance by

drawing a circle, to *mark* a direction by drawing a straight line, and to *indicate* a difference of two directions as the angle between two straight lines. Marking a distance or a direction is a primitive

measurement, reducing a quantitative relationship to perceptual

terms, the very purpose of measurement. Indicating a difference in

direction is, likewise, a primitive measurement, in the same way.

Euclid uses lines and circles to reduce measurement to its

perceptual base and to expand from that base.
Euclid's primaries, then, are the *measurements* of direction

and distance by *means* of circles, straight lines and, derivatively, angles.
Euclid's focus on measurement is not a focus on *numerical* measurement, but on directly perceptible, objective characteristics

of objects. For example, this distance is greater than that distance,

these two angles are equal, or, finally, this line intersects that circle.
Finally, notice that Euclid focuses on the

means of

measurement, not the *objects* of measurement. He focuses on lines and circles, his means, as opposed to the implicit objects of his

measurements: direction and distance. This focus on means,

without explicit reference to the object toward which these means

are directed, is an unfortunate tendency in Euclid. I will return to

this point in the next two chapters.

Euclid's Postulates and Common Notions

Euclid's

Elements is presented as a deductive system. He starts with certain basic axioms or, as he calls them, postulates and a second set of axioms, which he calls common notions.[72] The postulates all express something specific to the subject matter, geometry, while the common notions express more general truths. These axioms are not to be proven but are to be taken as true from the very beginning of the enquiry. Euclid, implicitly, held them to true and to be inherent in our observations of the world. But he left it up to the reader, based upon his understanding of the statements and reflection upon his own experiences to either assent to them or to abandon the enquiry as pursued by Euclid. These axioms should be regarded as fundamental observations about the subject of the enquiry. They formulate the perceptually given.

Having appealed to observation to support Euclid's short list of postulates and common notions, everything else in the subject, every geometric truth, is to be deduced from these axioms. Euclid organizes our geometric knowledge of the world by reducing one's reliance on direct experience to a few basic judgments and, in an orderly stepwise fashion, deduces everything else in the subject from these few basic judgments.

That so much can derive from so little is already remarkable and mysterious. But there is a greater mystery. For how

remarkable and mysterious. But there is a greater mystery. For how can such a system, based, it seems, on perfectly straight lines and exact equalities even apply to a world in which all of our measurements have finite precision; to a world in which no line is infinitely straight, infinitely circular, or infinitely thin? How can one ever say that two distances are ever equal if equality requires infinite precision? If straight line means infinitely straight, then there is no such thing as a straight line. So it can't be possible to draw a straight line connecting two points. And then Euclid's first postulate, along with all of his concepts, become floating abstractions, vindicating the Platonic interpretation that he himself may have held. Euclid's deductive system becomes a world onto itself.

If Euclid's postulates do not apply to the world, then his conclusions do not apply to the world either. So, to vindicate Euclid's propositions, one needs to start by identifying how his postulates actually do apply to the world, what they actually mean. One needs to understand what Euclid is implicitly saying about the lines and circles that actually exist on earth, taking into account the fact that all measurements have finite precision.

Yet, to reiterate, I do not claim that Euclid actually or consciously held the interpretation I offer herein. Rather I aim to

consciously held the interpretation I offer herein. Rather I aim to

make sense of Euclid insofar as he's saying something about the

world. I contend that, in the nature of the case, there is no other

earthly meaning to his work. And, finally, my interest is not in what

Euclid may have meant or understood about his own work, but on

what he *should* have meant. My interest is in what Euclid can teach us about the world.

In what follows, I will discuss each postulate in turn. I will

show in what way each of the five postulates formulates the

perceptually given, identifies and captures our ability to make

certain *perceptual* judgments that are essential to the measurement of distance or direction, and identifies certain basic, primitive

measurements, underpinnings of more complex measurements.

The postulates are about measurement, but none of them

presuppose or require numerical measurement. On the contrary,

they provide the foundation for indirect numerical measurement.

Finally, all of them are subject to contextual precision limits and are

universally valid within the appropriate contexts.

Euclid is famous for providing a deductive system. But

what makes it work is the fact that every argument in that system is

a chain of abstract measurements. As we shall see, regarding

Euclid's postulates as primitive measurements lays bare their

purpose and is the key to understanding and defending Euclid.

The Five Postulates

Euclid's five postulates read:[73]

1.

To draw a straight line from any point to any point.
2.

To produce a finite straight line continuously in a straight

line.
3.

To describe a circle with any center and distance.
4.

That all right angles are equal to each other.
5.

That, if a straight line falling on two straight lines make the

interior angles on the same side less than two right angles,

the two straight lines, if produced indefinitely, meet on

that side on which are the angles less than two right angles.

At first glance, these five postulates must seem enormously

puzzling. The first three are incomplete sentences. The fourth is a

complete sentence but its meaning and import is unclear. And the

fifth is simply a mouthful. Taken together they must seem a motley

assortment, an unlikely foundation for building a system of

geometry.

geometry.

Understanding the postulates as primitive measurements provides the key to understanding the meaning of each one taken separately and understanding the interrelationships among all of them. But, first, there is an important distinction separating the first three postulates from the last two.

Suppose a carpenter wants a two-foot board. He begins by measuring a number of boards. One is 20 inches, one is four feet, and a third one is 30 inches. When he ascertains the length of the third, for example, he has performed a measurement in the sense of *identification*. The board, prior to his measurement, was already 30 inches.

But the carpenter *wants* a 24-inch board. So he selects the 30-inch board and *measures out* 24 inches on it. In this, he is not determining the length of something; rather he is *finding* a position at a specified distance from one end of the board. Having marked the length he is looking for, he cuts the board at the point of his marking to obtain his two-foot board. Measuring out and then cutting off the excess is measurement in the sense of meeting a set of specifications or constructing something to a set of specifications.

Of the five postulates, postulates 1, 2, and 3 all call for constructions. Euclid indicates this grammatically by starting each of them with infinitives, such as “to draw” or “to describe”. On the other hand, postulates 4 and 5 are identifications of geometric relationships.

The proofs of Euclid’s propositions generally rely on a series of abstract measurements that include both types: both constructions and identifications. His constructions are the measuring-out steps. They provide the measuring *apparatus*, or set up the relationships, needed to support the subsequent identifications that establish the proposition.

Postulate 1: “To draw a straight line from any point to any point”[74]

The first postulate may be restated: A straight line can be drawn from any designated point to any other designated point. The postulate calls for a construction, but, in so doing, makes an assertion that the construction can be accomplished. And how might one do this, in practice? If the points are close together, one follows classical Greek practice: One uses a straight edge, first, to connect the points and, second, to draw the line indicated line. For points that are a somewhat greater distance apart, one stretches a string between them. For an even greater distance apart, one

a string between them. For an even greater distance apart, one typically lines them up by sight. One stands near one point and finds the direction for which the two points line up. For increased precision, and distance, one uses a scope. This process is usually referred to as finding a “line of sight”. Where these methods are unavailable, less direct methods are required, but methods of this sort are the foundation of the less direct methods. Whatever the method, it is important to grasp and take seriously that Euclid himself is not prescribing, or at least *need not* prescribe, *how* the construction should be accomplished or what physical form the line must take.

By common assent, Euclid actually meant, and, subsequently, requires, a slightly stronger statement, namely:

A straight line (and only one straight line) can be drawn from a point to any other designated point.[75]

Putting this particular qualification another way, two straight lines cannot, in Euclid’s geometry, enclose a space.[76] To see why this is important, consider that, on the earth, one could consider longitudinal lines running north and south as the closest available equivalent to straight lines on the face of the earth.

Indeed, on the perceptual level, and on the perceptual *scale*, longitudinal lines *are* straight, straight within the precision that the human eye can detect. Yet such lines intersect at both the North and South Poles and any two of them enclose an area on the Earth’s surface.

More generally, connecting any two points on the earth

More generally, connecting any two points on the earth, there is a “great circle” that represents the shortest, and straightest, path available on the Earth’s surface to connect the points. The more general mathematical term that embraces both straight lines and great circles is “geodesic”. A geodesic is a line that looks like, is indistinguishable from, a straight line on a sufficiently small scale, within a pre-specified precision limit. In general, a geodesic is the nearest equivalent to a straight line when that line is constrained to remain on a surface. A geodesic is a line that, on a sufficiently small scale, does not bend or curve.

In particular, a great circle can be characterized as the path that one takes when one continues to walk in what one takes to be a straight line. The first, weaker form, of Euclid’s first postulate is satisfied, in general, by geodesics, and, in particular, by great circles: There is at least one great circle on the earth connecting any two points. The stronger form, that the connecting geodesic be unique, is satisfied by straight lines, but is not satisfied by great circles.

This is where a discussion of the first postulate would normally end. We now understand what the first postulate means, how it relates to our experience, what Euclid actually had in mind, and how it applies to the Earth’s surface.

But my purpose requires a little more: What does Postulate

1 say about measurement?
We already know that a straight line determines direction.

To connect two points by a line, by whatever means, is to establish

and mark the direction of the second point from the vantage point

of the first.
Notice that one can *specify* a direction without drawing a

line, just by specifying two points. “From here to the moon,”

establishes a particular direction. But drawing a line (e.g., finding a

line of sight) *measures* the direction in the following sense: First, it provides us with a perceptual grasp of a particular direction: it

marks a direction. Secondly, it provides a way to find *other* points in same direction: any point on the line lies in the same direction

from the point of observation as any other point on that side of the

line from the point of observation. And, finally, it helps distinguish

points that lie in the specified direction from those that do not.
For example, a carpenter laying out the foundations of a

house will stretch a string tightly between two nails attached,

respectively to two stakes. Using the string as a guide, the carpenter

will drive in further stakes, as needed, at various points along the

string.
But does specifying two points really suffice to specify a

particular direction? Is there one and only one straight line that can

be drawn connecting the two points? Postulate 1 says the answer is

yes. So when I said that one can specify a direction simply by

specifying two points, I was relying on Postulate 1. In sum,

Postulate 1 says you can always find and mark a unique direction

from one point to another point; that any point lies in a

determinate direction from a point of observation. Euclid measures

that determinate direction by drawing the line that specifies the

direction between the two points.

Postulate 2: "To produce a finite straight line continuously in a straight line."[77]

Postulate 2 can be restated as: A straight line can be

extended, as needed, in either direction.
Whereas Postulate 1 says that one can find a direction

between two points; postulate 2 says that one can keep going in a

particular direction. Both are separate aspects of measuring

direction. Postulate 2 implies that finding a line of sight is *sufficient* for finding any required point in a specified direction from a point

of observation and distinguishing points that lie in that direction

from points that do not.
But why is this important? Why extend a line? Typically,

Euclid extends lines to intersect some other geometric object such

as a line or a circle. Because that intersection point is on the line, it

lies in the direction of the line. The intersection finds an object,

indeed a particular point on that object, lying in the specified

direction. Extending the line, in this fashion, is a measurement of the intersected object: It determines which part of that object lies in a particular direction.

Euclid relies on intersections constantly; they are the lifeblood of his entire system. Intersections are important because an intersection is a point that satisfies, simultaneously, two different conditions by dint of being part of two different geometric objects.

Euclid speaks of *drawing* a line because that is the way directions are indicated in his pictures. But the line he draws serves as a visual abstraction that covers a stretched string, the edge of a straight edge, a line of sight, or the straight path that one might take from one point to another. Once again, Euclid's concern is not, or need not be, with the particular means chosen in each circumstance, but with the fact that it can be done, that he can appeal to this ability within a geometric argument. Insofar as the line determined by one of these methods is regarded as straight, it is a straight line, i.e., covered by Euclid's abstraction.[78]

In this sense, when Euclid speaks of drawing a line, he is *specifying*, but not actually making the corresponding measurement. Euclid does provide pictures of the relationships he treats. But it's like writing a recipe for making something to eat.

The recipe specifies a *process* of mixing and cooking without actually doing the mixing and the cooking. In the same fashion,

Euclid specifies the measurement process that would be required to

establish the quantitative relationships asserted by his

Propositions.
To recap, Euclid's first postulate is taken to mean that only

one straight line can be drawn connecting any two points, that the second point lies in a specific and specifiable direction from the

first. The second postulate states that any direction, as specified by

a straight line, can be continued indefinitely.

Postulate 3: "To describe a circle with any center and distance."[79]

Postulate 3 can be restated: Given any center and a second

point, some distance from that center, a circle can be drawn

consisting of all points having that distance from the center.

Drawing a circle measures distance: It provides a way to

mark the distance, to *find other points* the same distance, but lying in a different direction, from the central point, and to distinguish

the points that, thus, lie on the circle from all other points that do

not.

This is the only direct way to compare distances that Euclid

wants to *assume*, as his starting point, to be possible. So his system does *not* rely on the assumption of tape measures or other portable standards. He *does not*

assume an ability to compare distances at different locations. Rather, he assumes *only* the ability to compare distances from a *fixed point*, regardless of direction.

How do we draw circles? For short distances, we use a

compass. So did the Greeks, although, unlike ours, their compasses

did not keep their radius after the circle was complete. One could,

in one act, find all the points of distance D, from a center C.

However, there was no direct way to draw *another* circle of that *same* radius D from a *second* center E. As for the first circle, one typically specified a distance D, by specifying a particular point

distinct from the center C and then using the circle to identify all of

the other points having the same distance from C.

Viewing the matter physically, it takes two points to specify

a distance. (Keep in mind that we are *not* talking here about numerical measures of distance.) For any distance so specified,

drawing a circle was Euclid's way of specifying all the other points

having the same distance from the point chosen as the center. It is

in this sense that Euclid uses circles and, implicitly, the compass to

provide his measure of distance or of length.

Because Euclid starts with the circle as his only means of

measuring distance, he will have to construct a means to compare a

distance between one set of points with a second distance between a

second set of points. Euclid will need a number of constructions to

do so. Indeed, following Euclid's statements of the postulates and

common notions, establishing the ability to compare lengths generally will be the first order of business in his first propositions. By the third proposition Euclid has established the ability to compare any two distances anywhere.

Postulate 4. “That all right angles are equal to each other.”[80]

Suppose someone on Mars tried to duplicate a yard stick. Without having some point of comparison, such as the distance, in yards, of Mars from the sun, this would be impossible. One needs a place to start. It would do no good to know that there are three feet in a yard and 12 inches in a foot, if one cannot determine the length of a yard.

But suppose someone on Mars wanted to duplicate a protractor. The only thing that one would need to know in order to build a protractor is how to subdivide the arc that contains the degrees. The angle that needs to be subdivided, depending on one’s perspective, is either the straight angle that forms the base of the protractor or, alternatively, half of a straight angle, namely a right angle. Once the Martian knew how many equal subdivisions to make, he could take his finished protractor to Earth and line it up against a protractor built on Earth; They would exactly match

against a protractor built on Earth. They would exactly match.

What is a right angle? Start with any straight line and

distinguish a point on it. Draw a second straight line, perpendicular

to the first through the point. That is, draw a straight line through

the selected point that makes the same angle with both sides of the

given line. *That* angle is a right angle.

Anyone anywhere can determine a right angle without

comparing it to any other right angle because a right angle is one

half of a straight angle. “Right angle” has an

independent

determinate meaning at any point in the universe.

Postulate 4 can be restated: The right angle, half a straight

angle, is the objective *standard* for measuring *differences* in direction. It can serve as that standard *because* right angles constructed independently in different places will match when they

are superimposed.

On the basis of Postulate 4,

recognizing a universal

standard, one can compare angles at different places. Because an

angle represents a difference in directions, Postulate 4 is the first

postulate about *differences* in directions.

Once the standard has been set, any other angle can be

measured or specified as a specific fraction or multiple of a right

angle.

Postulate 4 is needed because one cannot compare angles

on Mars to angles on Earth without recognizing or acknowledging

that a right angle on Mars is the *same* angle as a right angle on Earth. Later on, in Proposition 11, Euclid offers a construction to

create a right angle. By use of this construction, right angles are

constructed independently, though by the same recipe, at any two

points. Having been constructed independently, it is a separate

identification to state their equality. And it is also a separate *fact* that two angles, constructed in different places, but by the same

recipe, will match when brought into direct comparison. That

identification and that fact is Postulate 4.

Contrast Proposition 4 to Proposition 3. In the case of

length, Euclid provides a means of comparing distances in two

different directions from a central point. He will shortly deduce a

basis for comparing two distinct *lengths* at two different locations.

But, in the case of *angles* at different locations he takes a different tack, reflecting a unique fact about differences in direction. The

result is the fourth postulate.

Postulate 4 provides an objective standard for measuring

relative direction, *from a point.* But it does not provide a way to compare directions general, from different points. For example, to

compare the axes of rotation for Mars versus earth would require comparing directions at different points.

The fourth postulate is explicitly about measurement, offering the right angle as the basic standard for comparing angles.

What are Parallel Lines?

Euclid offers the following definition of parallel lines:

“Parallel straight lines are straight lines which, being in the same plane and being produced indefinitely in both directions, do not meet one another in either direction.”[81]

In contrast to the first four postulates, Euclid’s defining criterion for parallel lines does not appeal to something that one directly perceives; one does *not* perceive that two lines never meet. So, on the face of it, one has no reason to believe that parallel lines exist or can be identified as such. However, Euclid should be presumed to be aware of this gap and a reader should be looking for him to fill it. In effect, Euclid’s supposed definition is promissory; Euclid, in enunciating it, has promised to establish the existence of parallel lines.

Nonetheless, Euclid’s definition is not devoid of perceptual motivation even if Euclid does not provide such motivation. It is

reasonable to ask: What *do* we recognize perceptually? And this question has an answer. First, one observes perceptually that two lines can point in the same direction. One observes that the opposite sides of a rectangular table are pointing in the same direction, as opposed to two sides meeting at a corner, which point in different directions. Secondly when lines point in the same direction, one observes that they are neither converging nor diverging; we would not say that two converging or diverging lines are pointing in the same direction. Finally, while we may not be in a position to observe that two lines running in the same direction never intersect, one does observe the converse. One observes lines that do intersect. And when lines intersect, they point manifestly point in different directions.

Accordingly, two lines that *continue*, no matter how far they are extended, to point in the same direction will never intersect. This much, at least, is given in perception.

Despite one's justifiable first impression, there is a reason for Euclid's definition. Euclid, in fact, appeals as directly as possible to perception. Intersections are directly perceivable; measuring angles, for example, is not. For to compare two angles at different points requires a chain of comparisons, appealing ultimately to the

points requires a chain of comparisons, appealing ultimately to the fourth postulate.

Euclid will ultimately provide a criterion for two lines to be parallel and his criterion will consist in comparing angles.[82] But he is still one Postulate and 26 Propositions shy of being able to do that.

The perceptual meaning of parallel lines is: two lines that point in the same direction.[83] And one should take this concept to extend beyond immediate perception. For example, someone in the living room can be looking in the same direction as someone in the kitchen, whether or not they can see each other and whether or not anyone else is in a position to see the two of them.

I have stressed the fact that all measurements are subject to precision limits and this principle certainly applies to the first four postulates. All of Euclid's postulates are identified in a perceptual context and are subject to precision limits. However, recognizing precision limits and context is particularly critical in regards to parallel lines. To say that two specific straight lines, pointing in the same direction will *never* meet is to make an impossible claim. In particular, it is to say that one has measured the directions of the two lines with infinite precision, since the slightest difference will result in eventually meeting. Two straight lines pointing at the North Star are pointing in the same direction to an extremely high

degree of precision, but they will eventually meet at the North Star. Secondly, to say that the lines never meet is to say that the two lines are infinitely straight, since the slightest bending of either line would point them in different directions. Even the presumption that two lines are in the same plane is unrealistic when infinity becomes the standard. If two lines are ever so slightly skewed and fail to fall into the same plane, they point in different directions. Yet they will never meet; for if they did and were “perfectly” straight, they would lie in the same plane because, taken together, two intersecting lines determine a plane.

So, one cannot meaningfully take Euclid’s definition to require infinite precision; not if geometry is about the world. One needs to understand “same direction” as meaning “no relevant difference in direction”. Within any context one needs, at least implicitly, a threshold distance beyond which intersections are no longer relevant to parallelism, beyond which they do not correspond to a material difference in direction. Depending on the context, that threshold distance might be 100 feet, five miles, the distance to the North Star, or the distance to a distant galaxy. Euclid’s definition should be taken as a formulation of the fact that, on any scale and standard of precision for which two

distinct straight lines lie in the same plane and point in the same direction, they will never meet within the relevant threshold distance, within the threshold applicable to a particular context. Euclid, of course, does not acknowledge this limitation, but such an interpretation is implicit and is presupposed in any *application* of the concept, *parallel*, to the world. *Parallel* applies unambiguously within the context of a particular scale and precision standard. I will have more to say about this in Chapter 3.

Postulate 5: The "Parallel Postulate"

Postulate 5 is often referred to as the parallel postulate, even though it says nothing about parallel lines or their existence. But it does *complete* the required foundation to support Euclid's eventual criterion for two lines to be parallel. The formulation of a set of postulates providing such a foundation was a turning point in Greek geometry. Greek geometry has a long tradition going back to Thales and Pythagoras, but, as Aristotle attests,[84] geometers prior to Euclid were guilty of circular reasoning in their demonstrations concerning parallel lines. Euclid was the first to offer a postulate capable of breaking the vicious circle.[85]

Postulate 5 states: “That, if a straight line falling on two straight lines make the interior angles on the same side less than two right angles, the two straight lines, if produced indefinitely, meet on that side on which are the angles less than two right angles.”[86]

Postulate 5 states a criterion for two lines *not* being parallel, a criterion for their intersection, a criterion, therefore, for pointing in different directions.

Since the statement itself is a mouthful, the best way to understand it is to look at a picture. Accordingly, in Figure 3, X and Y are the two straight lines in question; the unnamed line is the line “falling on” the two straight lines. The interior angles are labeled A and C. I take the sum of these angles, A + C, to be less than two right angles (in modern terms, less than 180 degrees). Postulate 5 says that the two lines X and Y will eventually meet on the right.

In perceptual terms, the Postulate is saying that if two straight lines in a plane are pointing towards each other, are converging, then they will ultimately intersect.

Euclid states a condition on the interior angles A and C. But it seems more directly obvious to me that if angle B is greater than

angle A, then line Y is pointing toward line X on the right and that they will meet on the right. In any case, the two criteria are equivalent, as the argument in the diagram demonstrates.

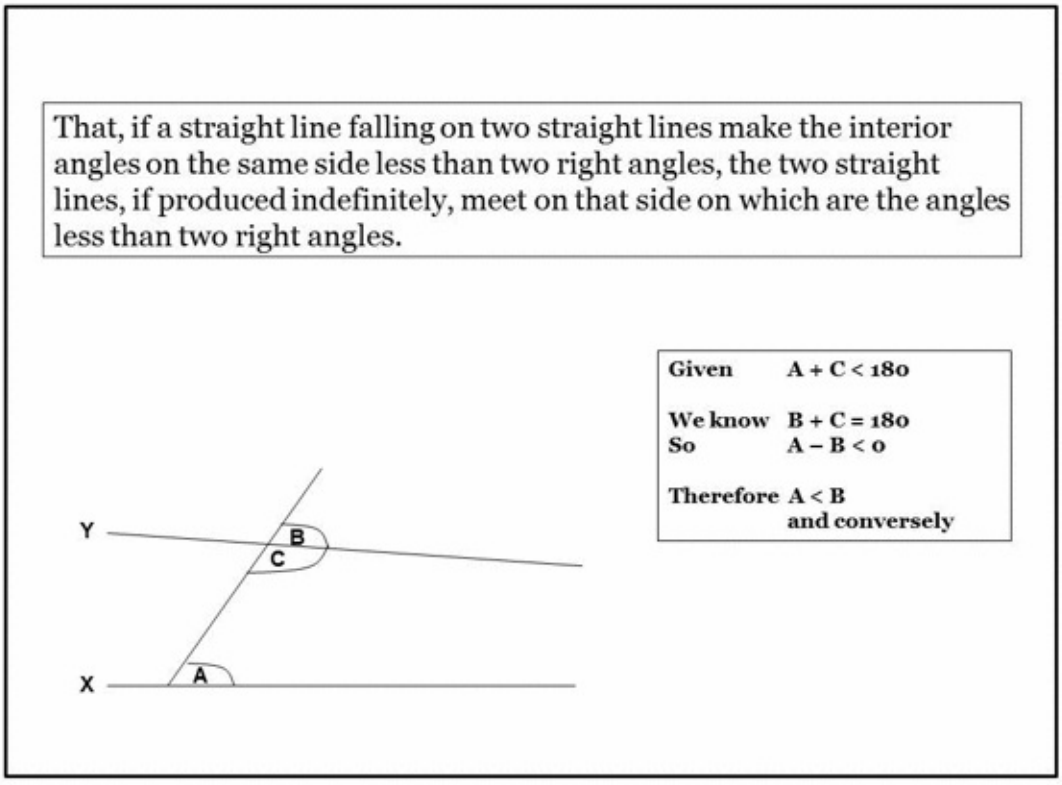


Figure 3

Recall that Euclid's Postulate 1 implicitly includes the statement that two lines can intersect in only one point – that there is a *unique* straight line connecting any two points. Postulate 5, by contrast, provides a criterion for when two lines *do* intersect. Euclid requires both Postulate 5 and the implicit portion of Postulate 1 to later adduce, in Propositions 27 and 29, a necessary and sufficient criterion for two straight lines in a plane not to intersect, that is to be parallel, that is to point in the same direction.[87] Namely, referring to Figure 3,

Proposition 27 says that if angle A equals

angle B, lines X and Y will not intersect, i.e., are parallel.

Proposition 29, in contrast, elaborates Postulate 5. In sum, X and Y

are parallel, i.e., will never meet, if and only if angle A = angle B.

Euclid's most important innovation was to find a set of

postulates sufficient to establish this criterion.
By this means, Postulate 5 provides a way to compare the

directions from two distinct points of observation. Two directions,

from two distinct points of observation, run in the *same* direction precisely when their respective lines of sight are parallel.
In this, Postulate 5 moves beyond Postulates 1, 2, and 4. It

says, in effect, that two converging lines will continue to converge,

ultimately intersecting

and thus manifesting a difference in

direction. The Parallel Postulate makes it possible to say that two

straight lines in different places are pointing in the same direction

or, alternatively, to say that those directions differ by an angle, say,

of 35 degrees. By virtue of the Parallel Postulate, one can compare

the directions of two straight lines in a plane by measuring the

angle that each line makes with a chosen reference line. And this is

the import of Propositions 27 and 29.
Postulate 5 and its consequences have sweeping

implications in Euclid's system. As a first example, the idea of

similar triangles, that shapes are scalable is a consequence that makes blueprints possible. If the Parallel Postulate did not hold, the angles of a triangle would change when one changed its scale; it would, for example, no longer be the case that the angles of an equilateral triangle would each be 60 degrees or independent of the size of the triangle. Secondly, Euclid's discussion and measurement of areas and, later, of volumes is entirely dependent upon the Parallel Postulate, as I shall trace in Chapter 3. Finally, astronomical measurements rely on the properties of parallel lines, even when such measurements require relativistic corrections. As an early example, Eratosthenes relied upon the properties of parallel lines to estimate the Earth's circumference.[88] More generally, astronomical measurement requires the use of trigonometry, the study of the relationships between the sides and the angles of triangles. But the development and application of trigonometry requires all five postulates as its base.

Modern textbook treatments of geometry do not generally follow Euclid's development. Instead, they typically jump directly to a formulation given by Playfair, namely that: "Through a given point one and only one line can be drawn parallel to a given straight

line."[89]
In perceptual terms, Playfair's version can be stated: At any

point there is a unique straight line pointing in the same direction

as a given straight line.
Also in perceptual terms, Euclid's Postulate 5 can be

restated: Two converging lines will ultimately intersect, manifesting

a difference in direction.
Both Euclid's and Playfair's postulates reflect

the

observation that lines in a plane that *don't* point in the same direction intersect each other. And, conversely, that lines that

intersect point in different directions at the point of intersection.

From either perspective, direction is the fundamental attribute

measured by Postulate 5.
There is a long history of failed attempts to prove Postulate

5 from the other postulates, an attempt that culminated in the

discovery in the nineteenth century that these attempts were all

doomed to failure. And this culmination included the development

of nonEuclidean geometries. Hitherto, across millennia, Euclidean

geometry had provided the recognized foundation of mathematics.

This discovery, and the eventual interpretation, of nonEuclidean

geometries occasioned the search for an alternative foundation of

mathematics.[90]
It is not my purpose to recount this history. However, my

general thesis does require further discussion of the status, the relevance, and the importance of the parallel postulate.

Accordingly, I return to this subject in Chapter 3.

In sum, Postulates 1, 2, 4, and 5 are related in the following way:

1. Postulate 1 says that one can find a direction between two points.
2. Postulate 2 says that one can find any point lying in a particular direction.
3. Postulate 4 says that one can compare *relative* directions (angles) measured at two different points.
4. Postulate 5 implies that one can compare directions, without qualification, between two different points.

These four postulates capture four separate aspects of what one recognizes perceptually in forming the concept of direction.

Perception is the Base

In the case of triangles, one's understanding of triangles is needed to understand more complex figures. The same principle

applies to the Postulates. For example, the study of geometry on the

surface of the earth, of the relationships between places and

distances on the earth, requires Euclidean geometry as its base.

This is also true for astronomical measurements and remains true

even insofar as measurement across astronomic distance is nonEuclidean. As I will elaborate in Chapter 3, Euclidean geometry, the geometry of the perceptual level, remains the frame of reference of

one's geometric measurements and provides the benchmark to

which all relativistic corrections must relate.

Both cases, then, exemplify the same principle: that all

conceptual knowledge must be related to the perceptually given.

Numerical measurement is meaningful *because* it specifies a quantitative relationship to something we can perceive.

Common Notions

Euclid distinguished five common notions,[91] namely:

1.

Things which are equal to the same thing are equal to each

other
2.

If equals be added to equals, the wholes are equal
3.

If equals be subtracted from equals, the remainders are

equal
4

4.

Things which coincide with one another are equal to one another

5.

The whole is greater than the parts

The first three of these constitute ways to deduce equalities from other equalities. Although these notions apply to quantity generally, Euclid's interest is to apply them to geometric properties. By contrast, the fourth common notion applies specifically to geometric figures. It is important because it provides an abstract statement of what we do when we compare objects by bringing them into close proximity. The final common notion provides a way to judge that one thing is greater than another thing.

All five common notions provide for making comparisons of equality or disparity, which is the essential underpinning of measurement.

Measurement and Euclid's Propositions

In reviewing some of Euclid's propositions, my focus will be on Euclid's *arguments* for those propositions. I am interested in the mathematical arguments insofar as understanding the mathematics helps identify Euclid's *method*. But Euclid's

propositions are also steps in his later arguments: each proposition, once demonstrated, becomes available to condense its argument for

later propositions. So, even as regards Euclid's method, one should

ask, for each of Euclid's propositions, "What does it accomplish?"

My discussion of the postulates showed the measurement

implications of each postulate, identifying, for each, the primitive

measurement it embodies. For Euclid's arguments the key question

is, "What is the role of measurement?" I ask this question not only

about the *arguments* for the propositions, but also about their *content*. Euclid's propositions are abstract statements of

quantitative relationships and are, themselves, abstract

measurements. A *series* of abstract measurements, taken as a whole, is an abstract measurement. To ask, and answer, what a

proposition accomplishes is to understand its measurement

implications.

Finally, the philosophical interest in Euclid's method is to

appraise the validity of Euclid's system, to understand just how the

statements of his propositions and the arguments for them apply to

the *world*. So the ultimate question for each proposition is, "Why are the proposition and its argument valid as applied to *real* shapes and geometric relationships?"

In discussing the postulates, I pointed out that Euclid's

measurements are not numerical, expressing quantitative

relationships to universal standards such as meters or kilograms

relationships to universal standards such as meters or kilograms.

They are more abstract than that. Every step in a Euclidian

argument either prescribes (constructs) or identifies a quantitative relationship. And every *proposition* either prescribes (constructs) or identifies a quantitative relationship.

A Euclidean argument is an extended chain of abstract

measurements that combines into an indirect measurement

establishing the quantitative relationship stated in the proposition.

But Euclid does not actually make these measurements. Rather, he

offers a *recipe,* a series of measurement instructions. A recipe, like a Euclidean argument, is a series of instructions that provides a

method to reach a desired result. In the case of a recipe for food,

one has to carry out the instructions and then taste the food to

appreciate the result. In contrast, to follow a Euclidean argument is

to understand the result of each step and, thereby, to apprehend the

outcome of the argument.

Finally, because the argument is abstract, both the

argument and its conclusion apply universally, in same way, to an

open-ended range of concretes. They apply whenever the set of

instructions, or any version of those instructions that has the

desired effect, can be performed to a degree of precision meeting

the requirements of the context. I will return to this point during

my discussion of Proposition 1.

Proposition 1. On a given finite straight line to construct an equilateral triangle.

Proposition 1 in Book 1 asks for a geometric construction.

Euclid prescribes a quantitative relationship and asks for a series of steps that will realize the prescribed relationship. Specifically, he asks for an equilateral triangle, a triangle for which all three sides are equal, starting from a given base.

Propositions that call for constructions follow a typical pattern, namely:

1.

Some information is regarded as being given.

2.

That information is specified geometrically.

3.

The Proposition requires the construction of a geometric object to the specifications provided by the given information.

Step three succeeds only if one can show that the specifications have been met. Success requires that each step in the construction constitute a valid measurement. It requires that the steps provide a series of quantitative comparisons between the

original geometric specification and the completed geometric object.

Proposition 1 reads: “On a given finite straight line to construct an equilateral triangle.”[92] The given, here, is the finite straight line. The other two sides of the finished triangle are required to have the same length as the given line. So those lengths will need to be measured out in the appropriate directions. To measure out lengths one needs to draw circles. To identify a direction from a point requires finding at least one other point in the required direction.

The construction is accomplished, in three steps, in Figure 4. One draws a circle around each endpoint with radius equal to the finite straight line. Each circle measures out the required lengths of the other two sides. The intersections of the two circles consist of those points that are simultaneously the required distance from each end point. Whichever intersection one selects finds the required direction in which the two additional sides of the triangle need to point and determines where these line segments need to end.

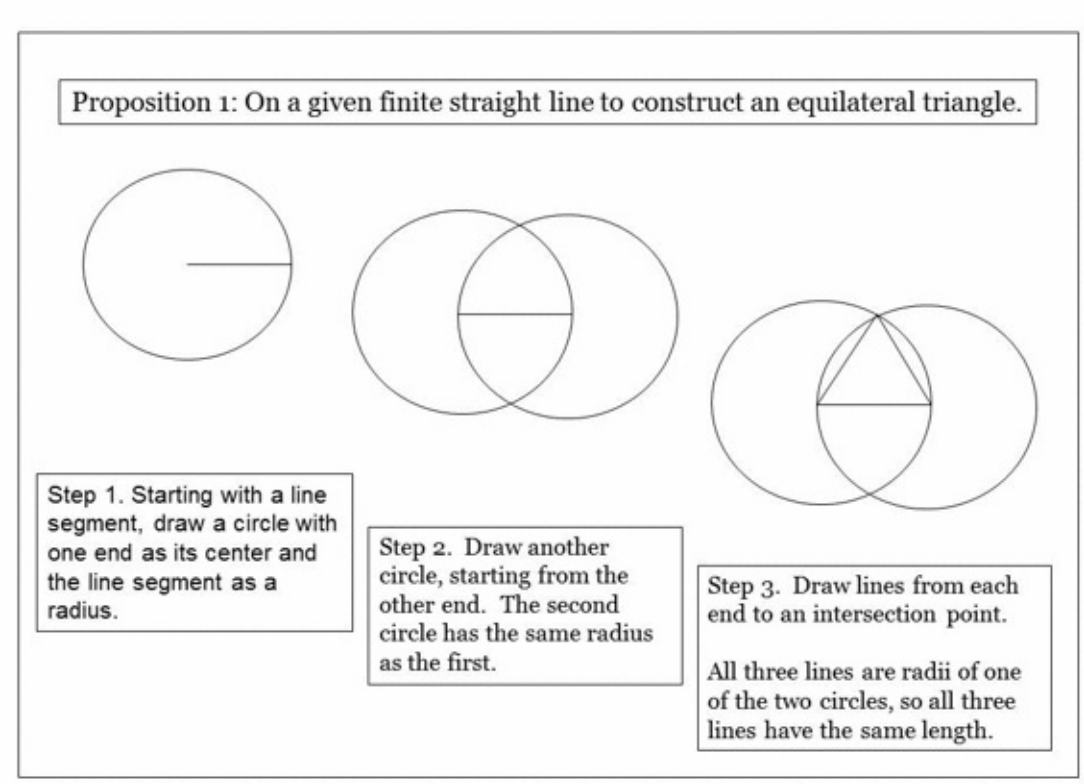


Figure 4

One says that the sides are equal “by construction,” referring to the act of measurement performed by the use of a compass to draw the circles. The sides are equal because one has used the compass to measure them. The sides can be compared because they have been measured.

Euclid’s measurement is an abstraction. It is not a concrete measurement of a single concrete instance, but a prescription or recipe that can be applied to any and every concrete instance. It applies to an open-ended range of concretes. Every such instance, within each specific context, will require a specific finite level of

precision. And Euclid's process, if carried out with sufficient precision, will produce an equilateral triangle within that required precision. It is not necessary to actually carry out these steps to know what they would achieve. And subsequently observing that the proposition applies to a particular case is an act of discovery in the following sense: The fact that the proposition applies does not depend on someone actually realizing that applicability. Euclid's process applies indifferently to each and every such physical situation. Euclid has specified a process of abstract measurement applying to actual and potential concretes that exist on earth, in the world that we inhabit.

In treating these measurements as totally precise, what Euclid is actually, perhaps unintentionally, implying is that they can be applied to every physical situation, regardless of the required precision level, provided that each step in the process is carried out with sufficient precision.

In applying Euclid's recipe, it is not necessary to actually draw circles, just to accomplish in some way what Euclid accomplishes by drawing his circles, i.e., finding the third vertex of the required triangle. For example, imagine that the remaining two sides have already been measured out and attached with hinges to

the two sides. One would then swing these remaining sides together until the two ends meet. Necessarily, the ends of the two sides will come together at the required third vertex of the equilateral triangle.

So what has been accomplished?

Euclid's full intent in proving Proposition 1 becomes clear in his proof of Proposition 2. Proposition 1 is a first step towards more general measurement of distance and constitutes a first building block toward that end. From this point on Euclid can simply ask for the construction of an equilateral triangle whenever it will help establish a quantitative relationship or provide a step in a required construction. Postulate 3 only provided a way to compare distances in different directions from a central point. Any comparability of distances or lengths beyond that point remained to be determined. Proposition 1 goes a step beyond Postulate 3. Considering the circle on the left, Proposition 1 finds a line segment that is not in that circle yet is known to have the same length as all the radii of that circle. This is a first small step toward being able to compare lengths generally. The importance of that first step will become clear in the construction of Proposition 2.

I have said that the construction called for by Proposition 1 is an abstract measurement. So is every step in the *argument* for Proposition 1.

Drawing the circles measures out the given line segment in every possible direction from each respective end point of that segment. Connecting those end points to one of the intersections finds the directions in which each of the required sides of the triangle must point. Identifying that the resulting triangle is, indeed, an equilateral triangle is, *ipso facto*, an identification of a quantitative relationship based upon one's understanding of what a circle is.

The argument applies to the world because indirect measurement applies to the world. A series of measurements is an indirect measurement. A series of measurements achieves a required finite degree of precision providing that the steps in that measurement are sufficiently precise. It applies to, and is valid in, any concrete instance for which the required precision is, in fact, achievable. The recipe, itself, is an abstract specification of what would be needed in any concrete instance. The specific precision requirement is left open, is what Ayn Rand refers to as an omitted measurement.[93]

The applicability of the proposition is subject to physical limits. Its context is determined by the applicability of the postulates

postulates.

The

applicability of the proposition is limited to

situations to which such a series of measurements would make

sense. The process offered in the argument, like any recipe, has

specific limits, but these physical limits are left open, not specified

in Euclid's statement or demonstration, left as omitted

measurements. Such limits pertain to the *scope* of its applicability, but not to its universal, open-ended applicability of the proposition

within that scope.

Euclid and Abstraction

It may be helpful to contrast the way that I look at Euclid's

treatment of Proposition 1 to the way that Plato would have looked

at it. First, we know what Plato has to say about the diagram that

Euclid uses: that the diagram is an imperfect representation, an

image, of something residing in his world of Forms.[94] What that something might be is less than clear because the diagram itself is

complex. Presumably there is a Form of an equilateral triangle, a

Form of a circle, and a Form of the line segment. But, are there a

separate Forms for each circle and separate Forms for each of the

two constructed line segments? And what about the entire network

of relationships embodied in the diagram? Does that have a Form.

as well? And what about the temporal character of the construction:

first do this; then do that, etc.? Plato's world of Forms has a

specifically a-temporal character.[95] While it may seem clear, at first glance and in a general way, what a Platonic mathematical universe

might mean, there is certainly a devil in the details.

By contrast, I maintain that the diagram is simply an

example to show how one might follow the recipe provided in Euclid's argument. The example helps one grasp the principles, the

specific steps in his argument/recipe. In effect one is expected to

view the example from an abstract perspective, as a concretization

of an

abstraction. The diagram helps one see how Euclid's

argument applies to each particular case.

What is the proposition about? Plato would say that the

argument pertains to some Universal Truth residing in his world of

Forms, a truth that is only imperfectly realized in the concretes on

this earth. Whereas I have maintained that the proposition is

specifically about those concretes; that the proposition applies

precisely to all of those concretes that fall within the required

precision limits set by each specific context within its scope.

Finally, Plato would have to regard Euclid's argument as

expressing a timeless relationship among ideas, among ideas residing in his world of Forms. Whereas I maintain that Euclid's argument is a recipe, a specification of a series of measurements, abstract measurements identifying specific quantitative relationships, that, when performed in sequence, establish the proposition. And I maintain that Euclid's arguments apply universally because they apply to each concrete case, regardless of the specific precision requirements of each case. And if Euclid does not specify either precision requirements or the specific means of carrying out his steps, the right way to regard these omissions, again in Ayn Rand's terms, is as omitted measurements.[96]

Use of an Intersection

The third vertex of the equilateral triangle, the intersection of two circles, is simultaneously the required distance from each end of the given line segment. The use of intersections is everywhere in Euclid; it is the lifeblood of his system. What, generally, is the role of intersections in Euclid?

Consider Figure 5:

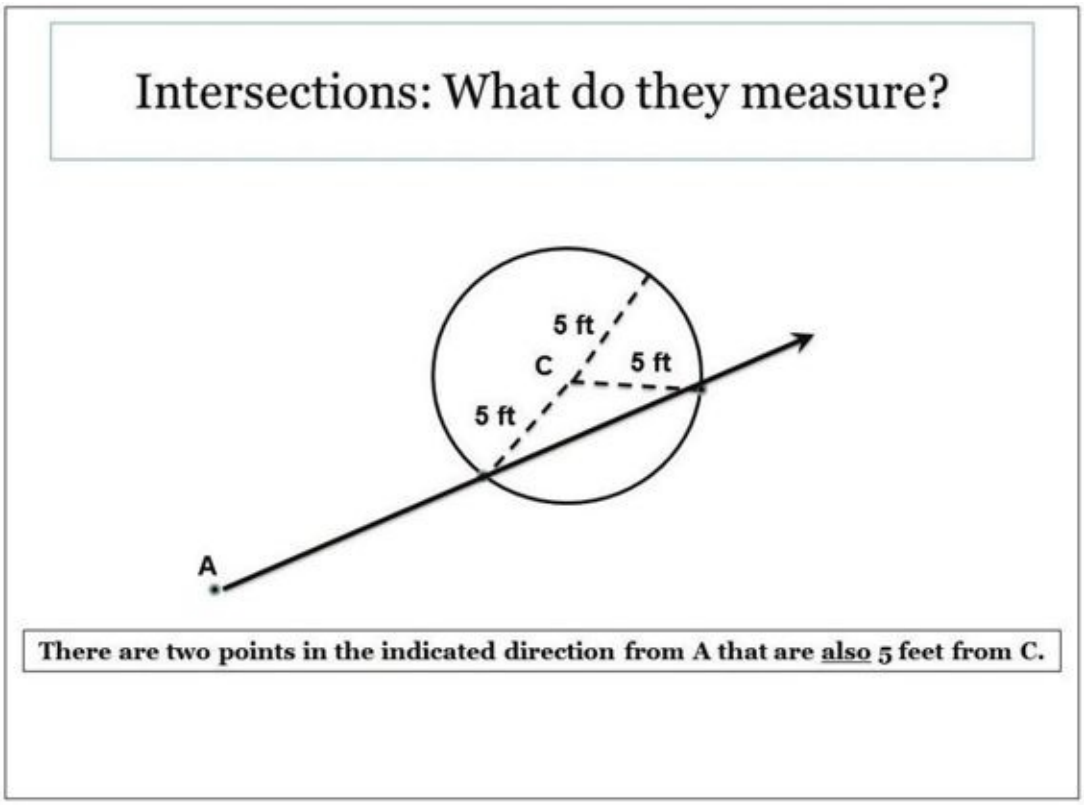


Figure 5

As Figure 5 makes clear, an intersection is a point that simultaneously satisfies two different measurements. In this particular example, the two intersection points are simultaneously points that lie in a specified direction from the point of observation labeled A and that are also a distance of 5 feet from the point labeled C. Finding a point of intersection is Euclid's equivalent to, is completely analogous to, solving two simultaneous algebraic equations.

Considering the importance of intersections in Euclid, it is important, though a commonplace, to point out that Euclid's set of postulates is not complete. For example, there are no postulates that cover the intersection of a circle and a line or the intersection of two circles. Euclid, without acknowledgement, appeals to perception of his figures whenever the need arises. While important, this lapse on Euclid's part in no way affects my broader point on the role of measurement or, specifically, of intersection, in Euclid.

Proposition 2. To construct a line segment at A equal to BC

Proposition 2 reads: "To place at a given point (as an extremity) a straight line equal to a given straight line."[97] Proposition 2 is a major step toward comparing distances or lengths of lines at different locations. The setup is shown in Figure 6.

Before presenting Euclid's construction, it is helpful to see what's really behind his construction. That is the specific purpose of Figure 6.

The first steps in Euclid's construction are what one might expect, given the construction in Proposition 1. Euclid begins by

connecting A to B and exploiting Proposition 1 to construct an equilateral triangle on the line AB. But how this will help is far from obvious. Somehow, one needs to leverage this equilateral triangle.

One can't physically move BC, but, as illustrated in Figure 6, one can do something very much like it. First, imagine that BC is connected by a hinge to B. Accordingly, rotate BC at B. Stop the rotation when it lines up with DB and lock the rotated line into alignment with line DB. Now imagine that the rotated line, together with attached line DB, is attached by a hinge at D. Rotate the entire line at D until it lines up with the line AD. In the course of this rotation, the point at B goes to A. So, in two steps, the original line BC has been moved to A.

That's the idea behind Euclid's proof. The equilateral triangle has provided the pivot point to rotate the required line segment to the point A. Euclid doesn't use hinges, but he uses circles to the same effect, taking his warrant from Postulate 3.

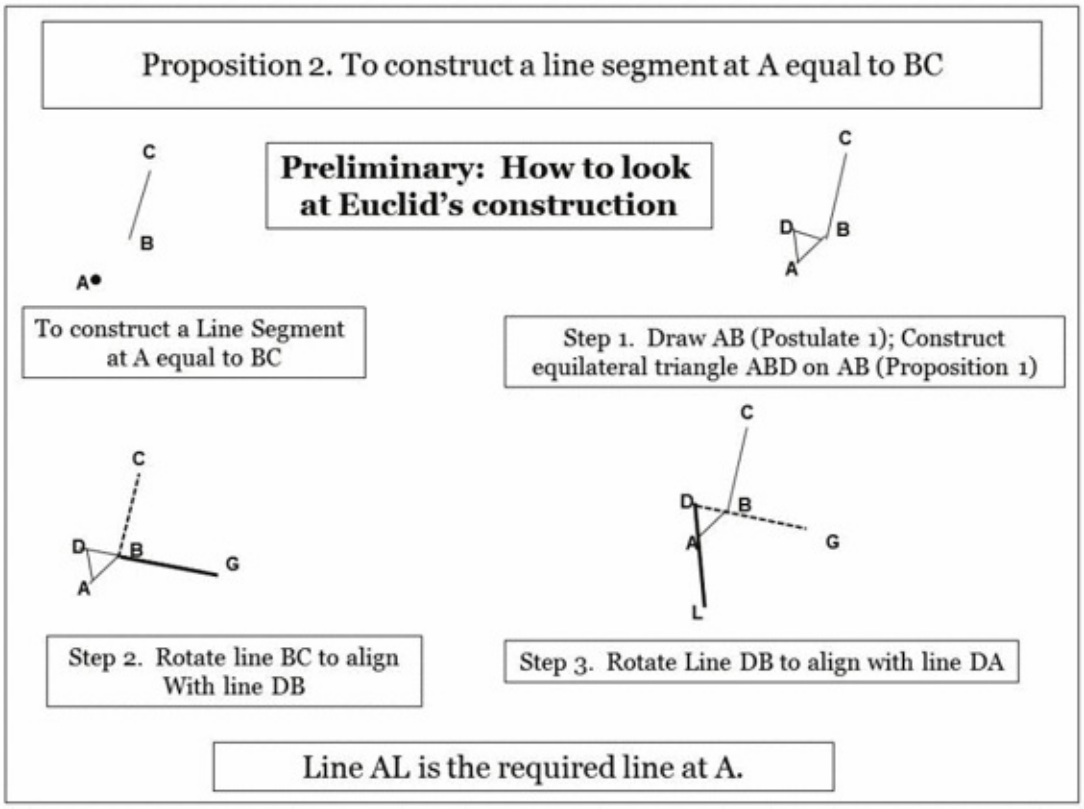


Figure 6

Euclid's actual proof is conceptually no different from the construction of Figure 6; Euclid draws a circle where I have rotated hinged line segments. And he finds an intersection where I have locked rotated line segments into place. The construction is depicted in Figure 7.

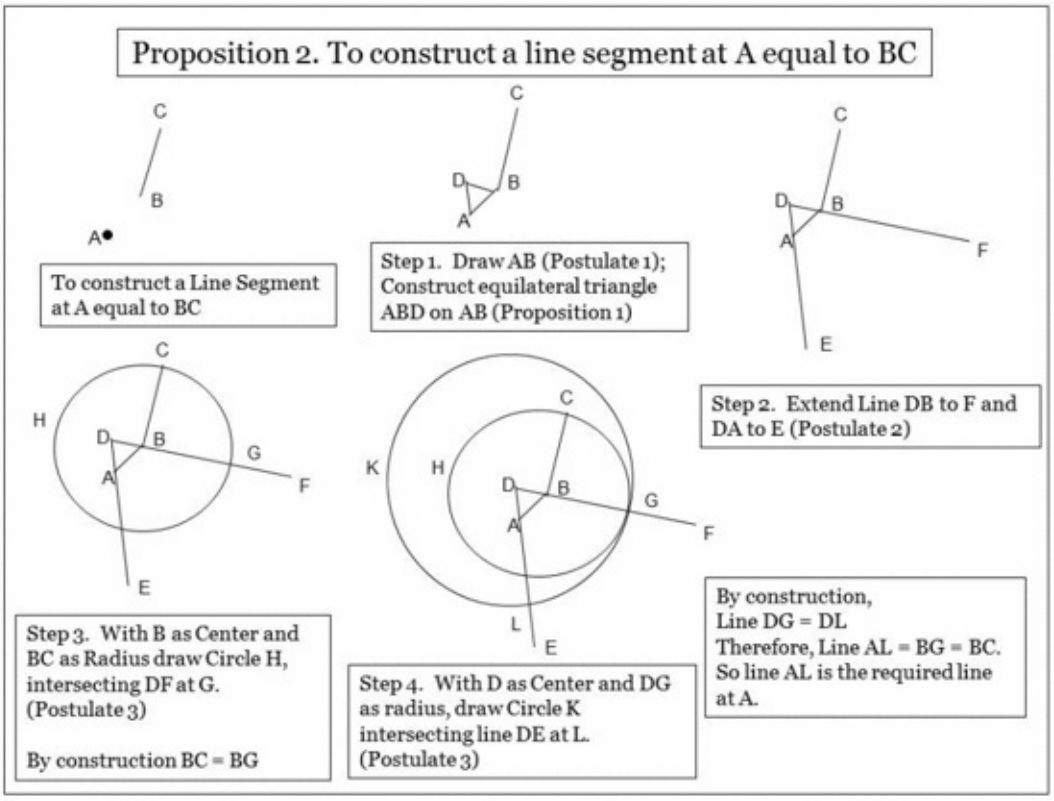


Figure 7

This entire construction should be viewed as a sequence of

abstract measurements, as follows:

1.

Joining A to B is finding the direction from A to B.

2.

Constructing the equilateral triangle (building on

Proposition 1) required finding a point simultaneously of

distance AB from both A and B. In the proof of Proposition

1, this required drawing two circles and finding their

intersection.

3.

Extending DA and DB required continuing a line in a particular direction, finding other points in the same direction.

4. Drawing a circle around B is finding a point on DF of distance BC from B.

5. Drawing a circle around D is finding a point on the line DE of distance DG from D.

To identify the fact that the construction succeeds, one needs to identify the nature of each step and appeal to known relationships. For example, the argument that AL = BC requires use of the common notions. AL = BG because DA = DB and because equals subtracted from equals (namely DL and DG) are equal. AL = BC because things that are equal to the same thing (namely BG) are equal to each other.

Proposition 3 reads: "Given two unequal lines, to cut off from the greater a straight line equal to the less."[98]

After Proposition 2, this Proposition requires one more step, namely to rotate the line that Proposition 2 constructs at A in the desired direction. (That desired direction is determined by a given line segment from the point A.)

So, by the end of Proposition 3, Euclid has shown how to

construct a line segment of any given length and direction at any chosen point. He has provided a recipe, shown an abstract way, to accomplish the same thing that we do whenever we move an object into a different position to compare it with another object. Taken together, the first three propositions specify, abstractly, how to move line segments around so that they can be compared.
The demonstration of Proposition 2 follows the exact pattern of Proposition 1. Its method is the same, its scope is the same, and its validity is the same. Together with Proposition 3, its corollary, Euclid can now appeal, at will, to the ability to compare line segments at any two separate points. The construction that he presents need never be explicitly repeated, but, simply, taken for granted in all subsequent arguments. Proposition 3 is a building block to be used over and over again. Whenever Euclid compares line segments, he is appealing to Proposition 3.

Proposition 4

“If two triangles have the two sides equal to two sides respectively, and have the angles measuring or comparing contained by the equal straight lines equal, they will also have the base equal to the base, the triangle will be equal to the triangle, and the remaining angles will be equal to the remaining angles

respectively, namely those which the equal sides subtend."[99]

Why can Euclid speak of two angles being equal? Euclid, though without acknowledgement, relies critically on the import of Postulate 4: Angles at different places are directly comparable because Postulate 4 says that the right angles at those different places are directly comparable.

Proposition 4 is ostensibly about triangles. But what it provides is a way of comparing angles, namely the two remaining angles that are subtended by the other sides. It also provides a comparison of the respective bases of the two triangles. More generally, Proposition 4 is the first step towards relating measurements of angles to measurements of length. This is the point of Proposition 4, its measurement implications.

What about its underpinnings? In the first three propositions, Euclid has established the basis for a direct comparison of two line segments. The quantitative comparisons to establish Proposition 4 proceeds as follows:

1.

First move the first side of one to coincide with the first side of the other. Because the two sides have the same

length, they coincide.

2.

The equal angles will also coincide because they are equal

and because they now start at the same place.

3.

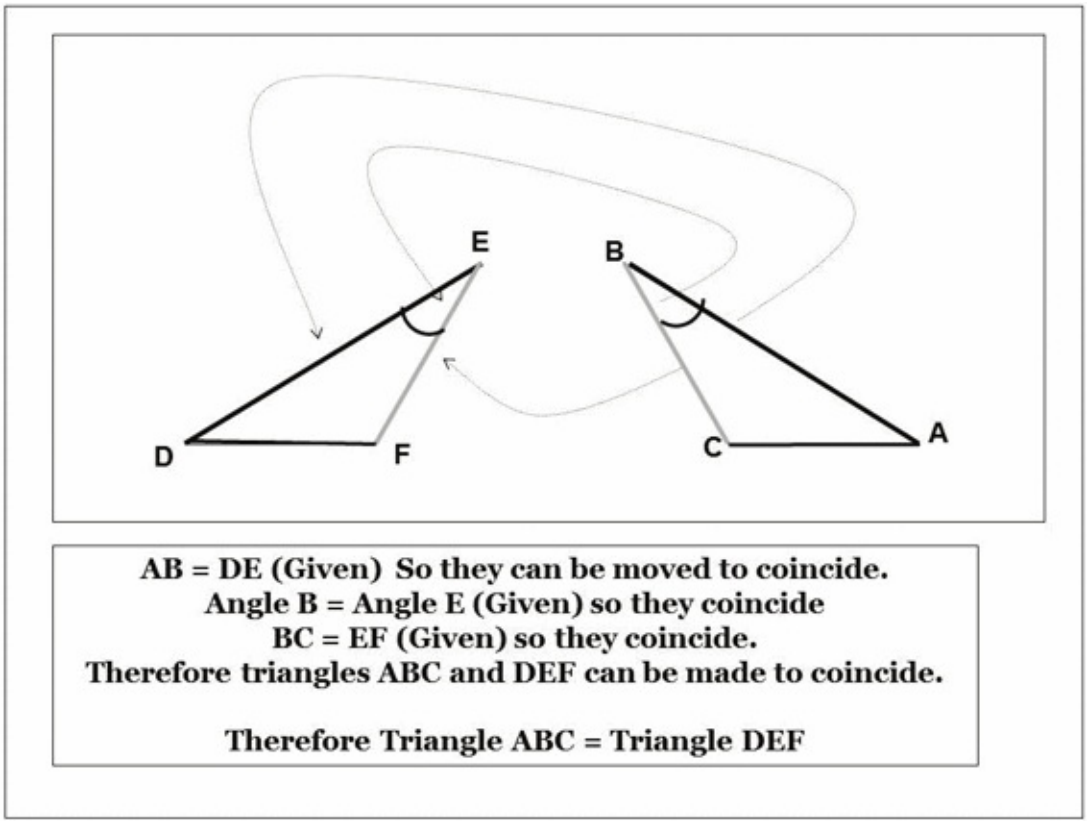
So the second sides will also coincide.

The entire argument appeals to the fourth Common Notion

that things that coincide are equal to each other. As Figure 8

illustrates, Euclid intends his argument to apply to reflections, as

well:



Proposition 4 tells us that two sides and the angle between

them determine the other parts of the triangle. But it does not tell us *quantitatively* how these three parts of the triangle determine the other parts. Euclid does not provide a formula. Yet this proposition is one of the foundations of trigonometry, later developed to do precisely that, to relate the measurements of the sides and angles of a triangle. And trigonometry, in turn, provides the underpinnings of navigation and astronomic measurement.

For Euclid, Proposition 4 is, in large part, a building block for further propositions. His characteristic use of Proposition 4 is already evident in the very next Proposition 5. Euclid typically will construct triangles that include the lines or angles that he wants to compare. Then he will argue that certain other parts of those triangles are equal. And then he will, finally, argue that the triangles are congruent and hence, the sides or angles that he's specifically interested in are equal.

Notice that, even by the time one reaches Proposition 4, making comparisons has gotten easier. The entire chain of measurements required to prove the first three propositions has now been consolidated into the statement of Proposition 3. The steps in these constructions need never be repeated in Euclid's later constructions. But, conversely, Proposition 4 is simply another landing point, one used constantly in Euclid's arguments for

subsequent propositions.

In sum, each proposition is a building block that facilitates further measurements. A geometric proposition provides the same kind of unit economy that Ayn Rand discusses in regards to concepts. In this respect, formulating a valid proposition, a geometric principle, is exactly analogous to forming a concept.[100]

Proposition 5. In isosceles triangles the angles at the base are equal …

Proposition 5 reads: “In isosceles triangles the angles at the base are equal to one another, and, if the equal straight lines be produced further, the angles under the base will be equal to one another.”[101]

Proposition 5 is not a construction; it calls for a comparison of two angles. But although it does not ask for a construction, Euclid used a construction to establish it. That construction is Euclid’s recipe to compare the two angles.

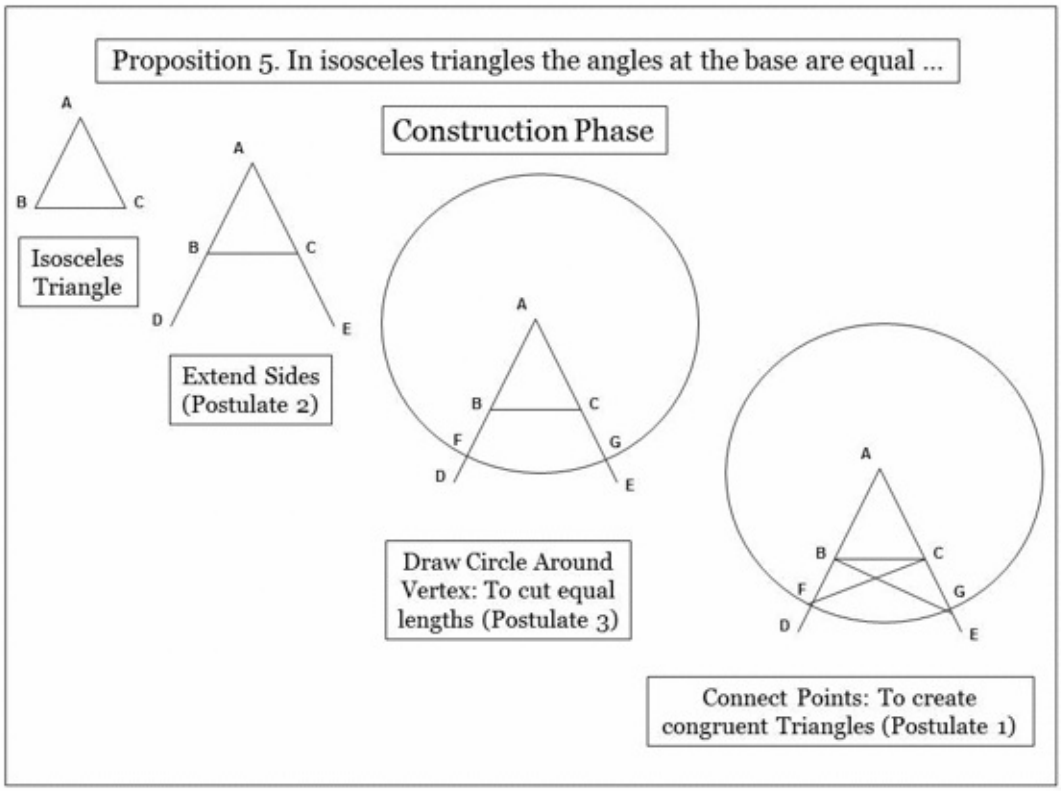
The steps in the construction are depicted in Figure 9. The first figure is the given isosceles triangle, which means that the two sides AB and AC are presumed equal. Step 1 is to extend the two

sides in their respective directions, ending in D and E. Step 2 measures off equal lengths AF and AG on these two sides by drawing a circle around A. The choice of a point F on the first line was a free choice but, once chosen, the intersection of the circle at G with the segment AE is determined by the radius of the circle. The final step is to connect C and F with the Line CF and to connect B and G with the line BG.

A number of triangles result from taking these steps. These triangles can be compared because of the knowledge one has already gained from earlier propositions.

Steps 1, 2, and 3 of the construction phase are all measurements. Step 1 extends the measurement of the directions specified by the two lines by extending the lines. Step 2 measures equal distances from A in the two directions. Step 3 measures the directions from C to F and B to G.

So ends the construction phase, depicted in Figure 9.



This construction, once it's complete, supports a series of

comparisons to establish the proposition. Focus, in Figure 10, on

the two large **bolded** intersecting triangles. I will not sketch a complete proof here. But, in outline, Euclid first compares triangles

ACF and ABG (the two large **bold** triangles in Step 1) arguing that they are congruent. "Congruent" means that the triangles can be

lined up so that they coincide so that all corresponding edges and

angles are equal. From that, Euclid argues, in Step 2, that the

triangles BCF and CBG (the smaller **bolded** intersecting triangles) are also congruent. Each step is a comparison, although each of

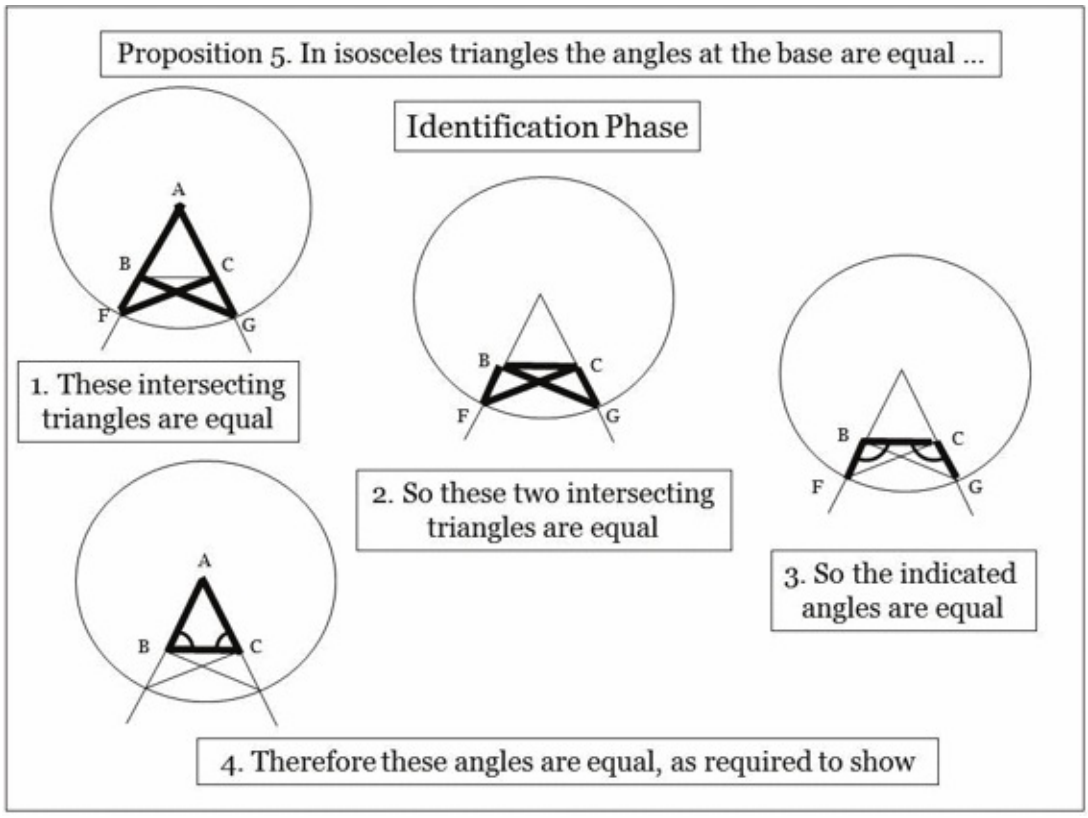
these comparisons could be broken down into further comparisons.

Even in this early proposition, Euclid draws on knowledge already

established in earlier propositions. Getting into these details, however, would neither affect nor further illuminate my argument.

In step 3, Euclid compares angles FBC and BCG (the marked angles), arguing that they are equal: a third comparison.

But, then, in step 4, the angles ABC and ACB (the marked angles) are also equal, a final comparison proving the Proposition.



I have omitted many steps, but this example indicates, in outline, the pattern of Euclid's proofs as generally including:

1.

A series of constructions, each constituting an abstract measurement and drawing on previously established knowledge, that is, on previously performed

measurements.

2.

A series of quantitative comparisons, each step the

ultimate product of direct quantitative identifications and

constructions.

Once again, Euclid has established his proposition by a

series of measurements.

As I have pointed out, Euclid's argument for Proposition 5

shows a typical use of Proposition 4: First, he constructs the

required triangles. Then he argues that certain line segments and

angles are equal. Next, he observes that these segments and angles

are parts of triangles and argues that the triangles are congruent.

Finally, he concludes that the remaining corresponding parts of

those triangles, including the parts of particular interest to his

Proposition, are equal, as well.

Other Consequences of Postulates 1–4

Proposition 6 is simply the converse of Proposition 5: If the

angles are equal then the sides are equal.[102]

The purpose of Proposition 7, which I will not state, is to

establish Proposition 8, which states that the angles of a triangle

are determined by the lengths of its three sides. Two triangles with the same

lengths of corresponding sides are congruent.[103] When one speaks of the rigidity of triangles, one speaks of Propositions 4

and 8.
Proposition 9 ("To bisect a given angle")[104] starts a new phase of Euclid's enquiry. Numbers enter for the first time. It was

not enough for Euclid or for the Greeks to speak abstractly of

dividing an angle into two equal parts or five equal parts. They

wanted an exact way to carry out the division. They wanted a

geometric construction. Indeed, they wanted a construction with

straight edge (the measure of direction) and compass (the measure

of length). Anything else was less than satisfactory. For example,

they could not trisect an angle with such tools and the quest did *not* end when Nicomedes discovered a way to trisect an angle with the

aid of a peculiar kind of curve called "conchoidal" lines, a curve that he was able to generate by means of a mechanical device.[105]
In establishing Proposition 9, Euclid bisects an angle as follows:[106]

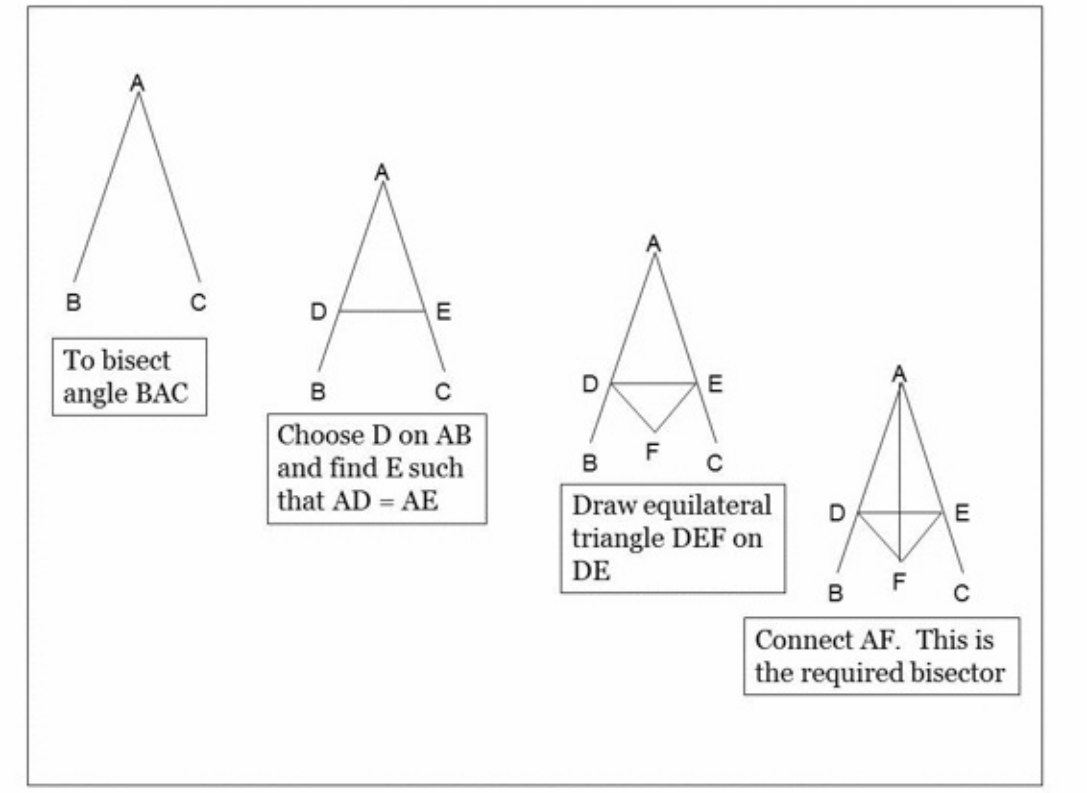


Figure 11

Proposition 9, in Euclid's treatment, already uses Proposition 8. It specifically uses the fact that the three angles of a triangle are determined once the three sides are known. Euclid argues that triangles ADF and AEF are congruent and that, therefore, the two angles DAF and FAE are equal, providing the required bisection.

Once again, one sees the application of an indirect measurement to establish an equality. In this case, a judgment of equality is used, for the first time, to establish a judgment of multiplicity: that the given angle has, indeed, been divided into

two equal parts.

The next three Propositions, 10 through 12 read:

●

Proposition 10: “To bisect a given finite straight line.”[107] ●

Proposition 11: “To draw a straight line at right angles to a given straight line from a given point on it.”[108]

●

Proposition 12: “To a given infinite straight line, from a given point which is not on it, to draw a perpendicular straight line.”[109]

These three propositions, together with Proposition 9, are all closely related. As in the proof of Proposition 9, Euclid’s constructions draw on earlier ones and, in this case, are not the most direct. In particular, he utilizes the results of Proposition 1. A more direct construction uses the concept that underlies Proposition 1 instead of simply invoking it. Euclid’s approach, which I will not discuss further, makes his demonstration quicker, but the underlying construction would take longer.

In Figure 12 below, the line CD is the perpendicular bisector of the line segment AB. This construction is the key to

Propositions 10 through 12:

Propositions 10 through 12.

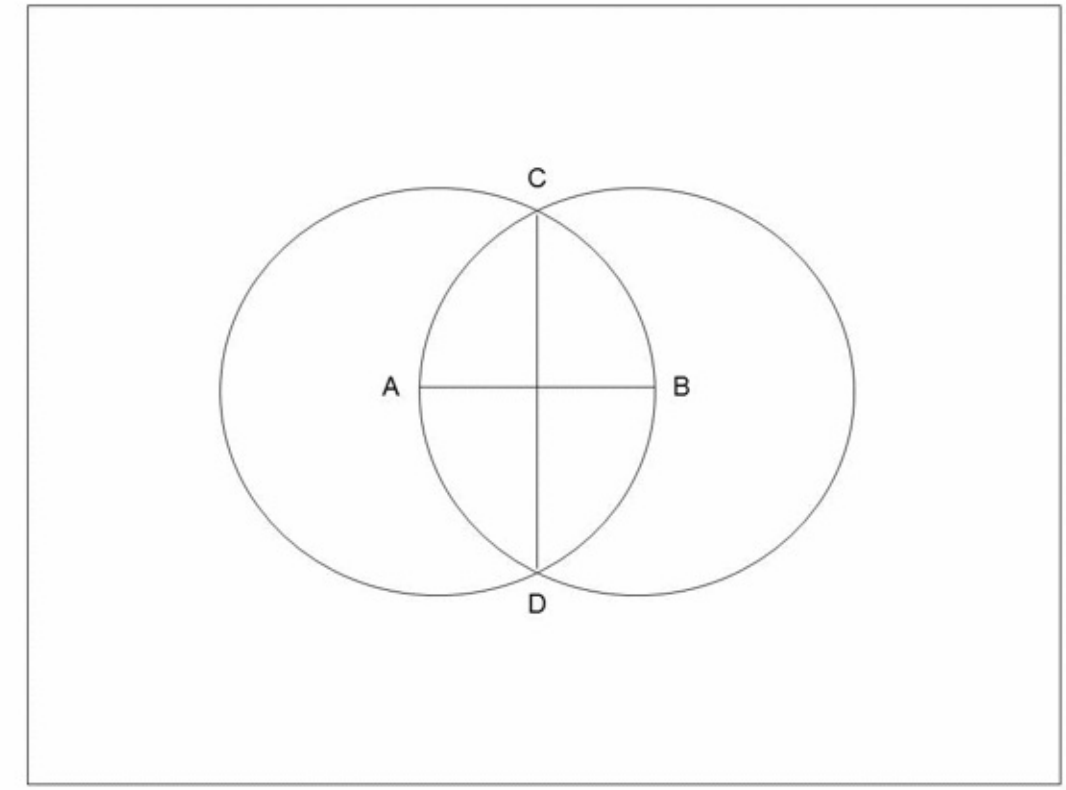


Figure 12

Taken together these Propositions are important because

1.

They provide the ability to subdivide.
2.

They provide the ability to construct a perpendicular to a

given line, either from a point on that line or from a point

not on that line. A line is perpendicular to another line

when it makes a right angle on either side.
3.

This *is* the recipe for creating right angles that I discussed earlier. As I discussed, the importance of Euclid's

Postulate 4 derives, in large part, from one's ability to manufacture a right angle at any point in the plane, or in the universe.

Proposition 26 is the third key proposition to establish conditions for congruent triangles. Recall that Proposition 4 says that a triangle is determined by two sides and the angle between them, while Proposition 8 says that the three sides of a triangle also determine the angles. Proposition 26 says that a triangle is determined by two of its angles and the line segment between them. It states: "If two triangles have the two angles equal to two angles respectively, and one side equal to one side, namely either the side adjoining the two equal angles, or that subtending one of the equal angles, they will also have the remaining sides equal to the remaining sides and the remaining angle to the remaining angle."[110]

Figure 13 outlines the essential points:

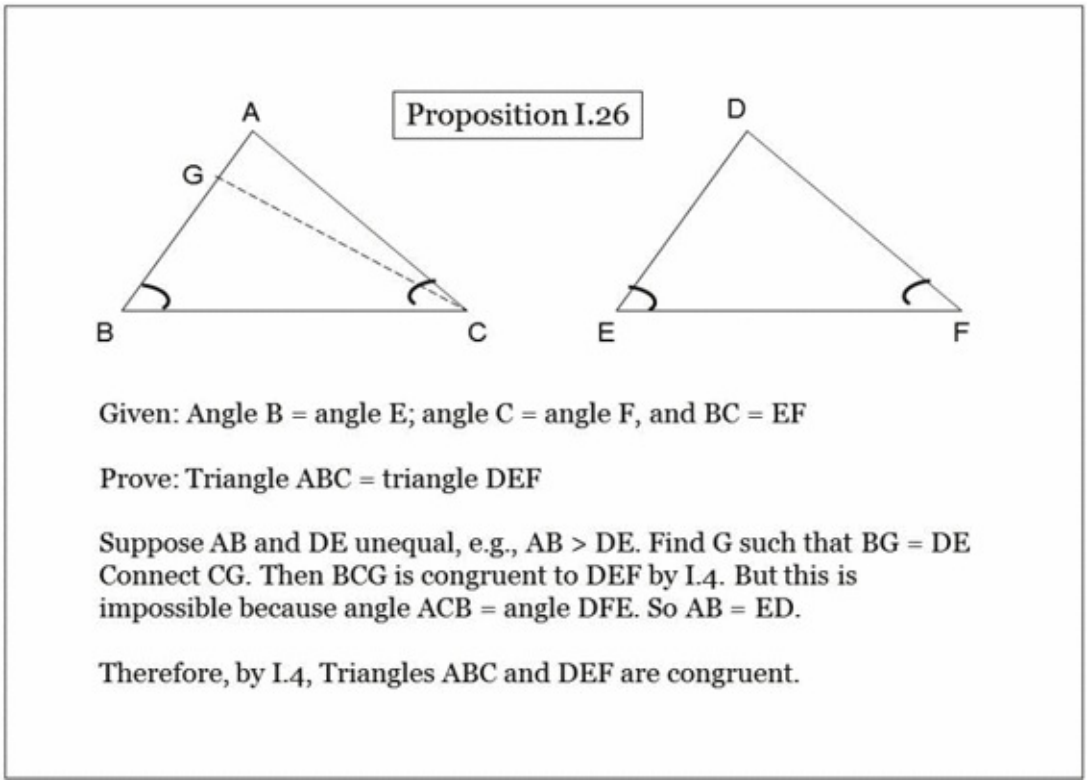


Figure 13

I will need this Proposition in Chapter 3 in relation to

Euclid's analysis of area.

I will continue my discussion of Euclid, in relation to

parallel lines, geometric area, and geometric proportion, in Chapter

3. But, by now, the role of abstract measurement in both the

content and the proofs of Euclid's Propositions should be evident.

The General Pattern and Conclusion

To summarize my review of Euclidean propositions:

1.

Every step in each of these demonstrations is a recipe for a

series of measurements.
2.

Each proposition condenses a series of measurements into

a single unit.
3.

Each step, in any concrete context, is subject to precision

requirements, to be met by a sufficiently accurate

execution of Euclid’s recipe.
4.

Therefore, each argument applies universally to every

circumstance embraced by the proposition.

A Euclidean argument is an abstract measurement of the world. Euclid does not know how precise any particular measurement will be or will need to be in any particular application. He does not need to know: his principles do not depend on any particular precision limit: They require only that there be a specific finite precision requirement in any particular application of his principles. Euclid leaves the problem of execution to those who would apply his principles. As applied to any specific context, he presumes:

- that the lines are thin enough and straight enough;
-

- that the measurements of direction are precise enough
- that the measurements of distance and the choice of the points from which those measurements start are accurate enough.

Euclid’s presumption is that his propositions will be applied to contexts for which the required precision is achievable by available physical means. The unknown inaccuracies can be ignored because they don’t materially affect the result. Euclid’s task is to proceed on that basis and to do nothing to introduce new inaccuracies in his own analysis.

Euclid’s arguments apply universally because they apply to each concrete case regardless of the specific precision requirements of any specific case.

The Need for Precision in Mathematics

If Euclid understood how his abstractions and arguments relate to the world, he certainly did not explain it. Yet despite the shortcomings in his presentation and in his understanding, his approach is fundamentally sound, once its proper relationship to the world has been identified and understood.

In particular, it is altogether appropriate, indeed necessary, to offer total precision in one's expression and analysis of mathematical relationships. The implication of such infinite precision, as I have argued throughout, is not that actual measurements ever attain such precision. To expect actual measurements to attain infinite precision is to miss the point. Any particular measurement will meet a finite precision requirement. But no specific precision requirement can be specified in advance, for all measurements, for all time. So the mathematical treatment can and must accommodate any and every required precision limit that will ever be encountered in any concrete application. From a reality-based perspective, this is precisely what is meant, what should be meant, by precision in mathematics.

Fundamentally, the need for precision in geometry arises from a difference between mathematics and engineering. In engineering, with a concrete application in mind, one can always identify the required precision in advance. The measurements one makes building a house are generally accurate within, perhaps, an eighth of an inch. Far greater precision is required for a precision machine and greater still is the precision required for semiconductors.

But the application of mathematics is open-ended. It applies to all engineering problems that will ever be tackled and all levels of precision that will ever be required. There is simply no way to anticipate the level of precision that may, someday for some reason, be required by someone for some purpose.

Euclid's abstract measurement meets the precision standard that mathematics requires. His arguments and propositions apply to all cases, whatever the specific precision limits of each case. His methods share the open-ended character that all concepts have; their applications extend beyond the scope of any special presumptions that one might make about the finest precision level that will ever be available or needed within any particular application. Propositions, and their arguments, apply, and are valid, to any context for which the required precision is physically possible.

Precision in mathematics means: independent of any a priori standard of precision.

In this way, mathematicians are able to achieve what an engineer cannot. A mathematician can analyze complex chains of mathematical relationships without ever losing precision. Mathematical methods are designed to provide whatever specific level of precision might be required without anyone having to

level of precision might be required without anyone having to know, in advance, the requirements for any concrete case.

Mathematical arguments are universal because their chains of mathematical relationships are independent of any specific finite precision requirement.

Mathematicians can indeed study and explicitly address questions regarding precision limits, but their arguments never depend upon any particular limit that one might specify in advance.

Mathematicians do make approximations in their work.

But when the approximation occurs within a mathematical analysis, a mathematician will always offer a way to meet any and all potential precision demands. The pattern is to say, and argue: for any positive number epsilon ($\varepsilon > 0$), there is a realizable approximation that is guaranteed to provide a final measurement within a range of $\pm\varepsilon$.

By contrast, any pre-specified approximation, set for all time, would fail to meet some precision level that might one day be required by someone for some purpose. The validity and universality of mathematical conclusions depends on the ability to analyze complex chains of mathematical relationships without ever losing precision.

What have we learned?

We have learned that geometric shapes are shapes that exist on earth. Concepts of geometric shapes are grasped perceptually and pertain to the shapes that we observe. Context is essential. Any specific context has its own standard of precision. Whether something is a triangle, circle, or straight line always depends on the standard of precision.

Secondly, we have learned that Euclid's propositions refer to shapes and relationships in the world. That a proposition about triangles applies to all triangles insofar as they are triangles.

We have learned that Euclid's postulates are the underpinnings of geometric measurement. That Euclid's Postulates formulate our most basic measurements on which the more complex measurements in the propositions are built.

And finally, we have learned that Euclid's arguments are valid, that they are recipes for measuring the world. A Euclidean argument is a recipe, a recipe Euclid argument as recipe prescribing a series of abstract measurements of the world.

[1] Euclid, *Elements*, edited with notes by Thomas L. Heath (New York: Dover Publications, 1956), Book I, Definitions 1 and 2

[2] Penelope Maddy, *Realism in Mathematics* (Oxford, Clarendon Paperbacks, 1992), For example, (p 28) "Let me return now to Platonism, the view that mathematics is an objective science." (p 28) Notwithstanding, Maddy later suggests (p 158) that her own view is closer to Aristotle. She continues to characterize her view as Platonic because "… it has become standard in the philosophy of

mathematics for any position that includes the objective existence of mathematical

entities." In any event, regarding the status of universals, Maddy, throughout her

book, presents some form of epistemological realism as the essential alternative to nominalism.

[3] W. V. Quine, "On What There Is," in Paul Benacerraf and Hilary Putnam, *Philosophy of Mathematics Selected Readings* (New Jersey, Prentice Hall, 1964

hardback). Quine states, (p 192) "Formalism, associated with the name of Hilbert …

the formalist keeps classical mathematics as a play of insignificant notations."

[4] Quine, (p 192)

[5] Haskell B. Curry, "Remarks on the definition and nature of mathematics" In Paul Benacerraf and Hilary Putnam. Curry states, (p 202) "According to realism,

mathematical propositions express the most general properties of our physical

environment … on account of the essential role played by infinity in mathematics, it is untenable today."

[6] Plato, *Republic*, (Plato 510d in standard numbering) "You … know how [geometers] make use of visible figures and discourse about them though what they

really have in mind is the originals of which these figures are images. They are not reasoning, for instance, about this particular square and diagonal which they have

drawn, but about the Square and the Diagonal; and so in all cases. The diagrams

they draw and the models they make are actual things, which may have their

shadows or images in water; but now they serve in their turn as image, while the

student is seeking to behold those realities which only thought can comprehend."

[7] Plato, Book 7 (527a) "[The geometers] talk of squaring and applying and adding and the like … whereas the real object of the entire subject [geometry] is …

knowledge … of what eternally exists, not of anything that comes to be this or that at some time and ceases to be."

[8] John Stuart Mill, *A System of Logic*, (London, Longmans, Green, and Co., 1919), Book II, Chapter V, "Of Demonstration and Necessary Truths" (p 154)

[9] Penelope Maddy, Realism in Mathematics (Oxford, Clarendon Paperbacks,

1992), (p 28), but see note 2

[10] Stewart Shapiro, *Thinking about mathematics* (Oxford, Oxford University Press, 2000 paperback), (p 55)

[11] Philip Kitcher, *The Nature of Mathematical Knowledge* (Oxford, Oxford University Press, 1984 paperback), (p 102)

[12] Ronald Calinger (editor) *Classics of Mathematics*, Prentice Hall, 1995. P 50 for Pythagoras dates, p 42 for Zeno, p 58 for Hippocrates, p 63 for Plato, p 74 for

Eudoxus, p 81 for Aristotle, p 109 for Euclid, p 131 for Archimedes

[13] Sir Thomas Heath, *A History of Greek Mathematics*, Volume 1, New York, Dover Publications, Inc., 1981, p 153-154 and p 167-168

[14] For a development of the Euclidean theory of proportion, see chapter 3

[15] Heath, p 154-155

[16] Carl B. Boyer, *History of Analytic Geometry*, Dover Publications, 2004. Boyer notes the influence of Zeno on p 7, "The crises which incommensurability caused in

Pythagorean philosophy and Greek mathematics might have been met by the

introduction of infinite processes and irrational numbers, but the paradoxes of Zeno blocked this path."

[17] Heath, p 275-278

[18] Heath, p 183-202

[19] Heath, p 322-335

[20] Euclid, especially book V, including Heath's notes

[21] I discuss Eudoxus's theory of ratio in Chapter 2, its application to similar triangles in Chapter 3 and the modern approach to irrational numbers in Chapter 4

[22] Richard Dedekind, *Essays on the Theory of Numbers*, Dover Publications 1963 from a 1901 English translation

[23] Archimedes, *The Works of Archimedes*, 1897, Cambridge: at the University Press, "On the Sphere and Cylinder" p 1-90, "Measurement of a Circle" p 91-8, and

"On Conoids and Spheroids" pp 99-150.

[24] Archimedes, "Measurement of a Circle" and "On Conoids and Spheroids"

[25] Harry Binswanger, "Selected Topics in the Philosophy of Science", 1987, on this characterization of straight lines, Available from the Ayn Rand Bookstore:

www.aynrandbookstore.com. My treatment of geometry is particularly influenced by

Binswanger's identification of length and direction as the primary concepts in

geometry. Binswanger takes *direction* as the primary concept in geometry, based on "to" and "from" the perceiver. He characterizes, a straight line as, in essence, a line of constant direction (vs. the changing direction of a curve and the discontinuous

direction of a jagged line), and parallel lines as straight lines in the same direction.

[26] Ayn Rand, *Introduction to Objectivist Epistemology Expanded Second Edition* (New York: Meridian, April 1979), p 42-47, p 193-194, on the role of context

[27] Rand, p 10-13, "If a child considers a match, a pencil, and a stick, he observes that length is the attribute they have in common, but their specific lengths differ.

The *difference is one of measurement*. In order to form the concept "length," the child's mind retains the attribute and omits its particular measurements."

[28] Rand, p 12, "… the term "measurements omitted" does not mean, in this context, that measurements are regarded as non-existent; it means that

measurements exist, but are not specified. That measurements *must* exist is an essential part of the process.

The principle is: the relevant measurements must exist in *some* quantity, but may exist in *any* quantity."

[29] Rand, p 193-194

[30] Harry Binswanger, "The Possible Dream" in *The Objectivist Forum*, New York, TOS Publications, Inc. April 1981, p 5, available from the Ayn Rand Bookstore, "What one sees under a microscope is not relevant to the concept of "sphere," which is formed to denote precisely the sort of shape possessed by billiard balls. A billiard ball *is* a perfect sphere, by a rational, contextual standard. A "perfect sphere" means a sphere that is flawless in the context of man's form of perception."

[31] Rand, pp 10-13 and 17-18, compare notes A27 and A28

[32] Ayn Rand, Introduction to Objectivist Epistemology Expanded Second Edition, April 1979, p 11

[33] Rand, *Introduction*, p 27

[34] Rand, *Introduction*, p 12

[35] Mill, p 147

[36] Mill, p 154

[37] Mill, p 149

[38] Mill, p 154

[39] Rand, p 10-13 and p 17-18

[40] Rand, p 7

[41] Plato, *Republic*, 510d

[42] Hippocrates George Apostle, *Aristotle's Philosophy of Mathematics*, Chicago, University of Chicago Press, 1952, p 3-4, Aristotle does not say this explicitly, but the pieces are there. Apostle reaches this conclusion by comparing Aristotle's discussion of quantity in the *Categories* (4b20-25) with a more specific discussion of the subject matter of mathematics in the Metaphysics (1061b17-25 and 1077b12-22).

[43] Jeremy Gray, *Plato's Ghost the modernist Transformation of Mathematics*, Princeton University Press, 2008 For an account of this transition, see especially, Chapter 3, p 113-175

[44] Moritz Epple, Chapter 10 "The End of the Science of Quantity: Foundations of Analysis, 1860 – 1910" In *A History of Analysis*, edited by Hans Niels Jahnke (Rhode Island, American Mathematical Society, 2003 hardback) p 291

[45] Apostle, p 4-8

[46] Quine, p 192

[47] Alfred North Whitehead and Bertrand Russell, Principia *Mathematica*, 3 vols, Cambridge: Cambridge University Press, 1910, 1912, 1913

[48] Jean van Heijenoort, 1967, *From Frege to Godel: A Source Book in Mathematical Logic, 1879-1931*, Harvard Univ. Press, See also, Mary Tiles, *The Philosophy of Set Theory*, Dover Publications, 2004

[49] Georg Cantor, "Letter to Dedekind" (1899) reprinted in Jean van Heijenoort, p 113, See also, Joseph Warren Dauben, *Georg Cantor His Mathematics and Philosophy of the Infinite*, Princeton, Princeton University Press, 1990, Chapter 6, "Cantor's Philosophy of the Infinite" especially, pp 124-8

[50] Maddy, p 28

[51] Curry, p 202

[52] Michael D. Resnik, *Mathematics as a Science of Patterns* (Oxford, Oxford University Press, 2004 paperback), p 4

[53] Resnik, p 201

[54] Rand, p 7

[55] Rand, p 8

[56] Rand, p 7

[57] Ronald Calinger, “Eratosthenes” in *Classics of Mathematics*, edited by Ronald Calinger (New Jersey: Prentice Hall, 1995), p 153

[58] I am not, herein, concerned to precisely isolate a category of *direct* measurement. Certainly, direct measurements all involve perceptual identifications.

But one can, for example, regard counting as a complex activity, yet it would not

defeat my purpose to place such perceptual judgements into the direct measurement

category. My account of the importance of indirect measurement does not hinge on

where one draws the line.

[59] Heath, p 153-154 and p 167-168

[60] Euclid, Book X, Proposition 5, including Heath’s notes

[61] Euclid, Book VI, Proposition 4

[62] See Euclid, Book VI, Definition 1 of “similar rectilineal figures” is Euclid’s term, which presupposes or anticipates Book VI, Proposition 4. On the contrary, I

use the term “similar” in its usual sense to call out the fact that the corresponding angles are the same. Euclid’s statement of Proposition 4 uses the term “equiangular”

instead of “similar”

[63] Rand, p 40

[64] Euclid, Definition 20

[65] Euclid, Proposition 6

[66] Binswanger, “Selected Topics,” See note 25, Binswanger considers this the essential characteristic of a straight line

[67] Euclid, Book I, Postulate 2

[68] Euclid, Book I, Postulate 1

[69] This is the meaning of Euclid’s use of the word “direction” in Book I, Definition 23

[70] Euclid, Book I, Heath’s note on Definitions 8, 9, “Definitions of angle classified,” Heath is most critical of the kind of view I’ve taken here. Most tellingly, he says, “Direction is a *singular* entity: there cannot be different sorts or degrees of direction.”

[71] See previous note. On this point, Heath’s characterization of direction as a singular entity is entirely applicable, is a way of observing that direction is not a magnitude.

[72] Euclid, Book I, Collected as the Postulates and the Common Notions, immediately following the Definitions

[73] Euclid, Book I, the Postulates

[74] Euclid, Book I, Postulate 1

[75] Euclid, Book I, Heath discusses this in his notes to Postulate 1. This interpretation of Postulate 1 is

required in the proof of Proposition 16 which, in turn, is part of the basis of Proposition 27

[76] Euclid, Book I, Heath notes to Postulate 1

[77] Euclid, Book I, Postulate 2

[78] How straight must it be? As in all such cases, the required precision and the available precision both depend on the context. In this regard, see Ayn Rand, p 196, "… you cannot go into infinity, but in the finite you can always be absolutely precise simply by saying, for instance: "its length is no less than one millimeter and no more than two millimeters."

[79] Euclid, Book I, Postulate 3

[80] Euclid, Book I, Postulate 4

[81] Euclid, Book I, Definition 23

[82] Euclid, Book I, Proposition 27 and 29

[83] Binswanger, "Selected Topics" See note 25. As noted there, Binswanger characterizes parallel lines as straight lines in the same direction

[84] Heath, *Mathematics in Aristotle*, Bristol, Thoemmes Press, 1949, Thomas Heath points to a passage in Aristotle's Prior Analytics (65a4-9), and concludes,

"This difficult passage, …, seems to imply that the theory of parallels current in

Aristotle's time involved some *petitio principii*."

[85] See also Heath's extensive commentary on Euclid, Book I, Postulate 5

[86] Euclid, Book I, Postulate 5

[87] Euclid, Book I, Proposition 27 and 29

[88] Calinger, "Eratosthenes"

[89] Heath, in his commentary on Euclid, Book I, Postulate 5, points out that this formulation was known to Proclus

[90] Marvin Jay Greenberg, *Euclidean and Non-Euclidean Geometries Development and History*, New York, W. H. Freeman and Company, 1983, Chapter

6, "The Discovery of Non-Euclidean Geometry", '' 177 – 222. See also Jeremy Gray,

Worlds Out of Nothing A Course in the History of Geometry in the 19th Century, Springer, 2007

[91] Euclid, Book I, Common Notions

[92] Euclid, Book I, Proposition 1

[93] Rand, p 12

[94] Plato, 510d

[95] Plato, 527a "[The geometers] talk of squaring and applying and adding and the like … whereas the real object of the entire subject [geometry] is … knowledge … of what eternally exists, not of anything that comes to be this or that at some time and ceases to be."

[96] Rand, p 12

[97] Euclid, Book I, Proposition 2

[98] Euclid, Book I, Proposition 3

[99] Euclid, Book I, Proposition 4

[100] Rand, Chapter 7, “The Cognitive Role of Concepts”, p 62-74

[101] Euclid, Book I, Proposition 5

[102] Euclid, Book I, Proposition 6

[103] Euclid, Book I, Proposition 8

[104] Euclid, Book I, Proposition 9

[105] Sir Thomas Heath, *A Manual of Greek Mathematics*, Dover Publications 1963, Ch VII, p 150 in a section entitled “The conchoids of Nicomedes,” The

conchoid is a curve, constructed by a mechanical device (around 270 B.C.) for the

specific purpose of solving this sort of problem

[106] Note that this construction does not work for straight angles. So constructing a right angle will require a different construction

[107] Euclid, Book I, Proposition 10

[108] Euclid, Book I, Proposition 11

[109] Euclid, Book I, Proposition 12

[110] Euclid, Book I, Proposition 26

Chapter 2 Measurement and the Geometry of Magnitudes

The most profound ideas in mathematics make their first appearance in arithmetic. The break between mathematics and reality in people's minds, though unnoticed, begins there, as well. The break is insidious *because* it goes unnoticed. To fill a gap one must first know that it is there.

If one wants to appreciate the subtleties of higher mathematics, one needs to appreciate their first appearance in elementary mathematics. If one wants to understand the tremendous mathematical advances in the nineteenth and twentieth centuries, to understand what they say about the world we live in, how mathematics relates to that world, how mathematics, at all levels, illuminates and quantifies our grasp of that world, one needs to begin with elementary mathematics.

By "understand" I mean to acquire a firm grasp of the relationship of mathematics to reality. So I am not speaking of conventional mathematical definitions and proofs by which mathematicians deduce the fundamental relationships in arithmetic from a few basic assumptions.[1] Nor am I speaking of the ability to compute or to manipulate algebraic expressions. And, finally, I am not speaking of the ability to use arithmetic in the everyday sense of balancing ones checkbook or calculating percentages.

Rather, I speak of taking a step back to understand just how numbers are used to measure relationships in our world, to isolate just what is the relationship of these numbers to the quantities that they are used to measure.

My specific concern will not be with counting objects, but with using numbers to measure magnitudes, such as length, weight, and speed. In this, we should not be surprised to find that our usage of numbers is indeed correct. But we will find that characterizing exactly what we are doing when we apply numbers is not as straightforward as one might have thought. Yet in laying this process bare, one creates the foundation for a similar understanding of mathematical concepts whose relationship to the world we live in may be far from obvious. It is the lack of such understanding that has led to the widespread false alternatives that mathematics is either a formal game played with symbols, a system of deduction

mathematics is either a formal game played with symbols, a system of deduction from carefully chosen axioms such as the axioms of set theory, or an insight into a Platonic universe of mathematical concepts. On any of these views, the applicability of mathematics to reality must be viewed as a happy accident.

We are taught to think of numbers as points on a real number line. There is value in such a perspective, but there is danger, as well. The danger lies in what is being ignored or taken for granted. The real number line is a culmination of a long historical development. It had to be long, because it is the product of many layers of abstraction. The well-known reluctance, at every step throughout history, to recognize ratios, negative numbers, and irrational numbers as bona fide numbers was not accidental and reflects the shifting perspectives required to embrace each newcomer.

A culmination of this development, the real number line, is not the place to start; not if one's goal is to fully understand the relationship of numbers to reality. It is one thing to reach such an abstraction, to clearly know its lineage, to see how each step ties to the previous and how each step applies to and illuminates the concretes of the world we live in. But it is quite another to begin there, as though its application to reality were self-evident and no further understanding were needed or possible.

Exploring the lineage of the real numbers involves history because the issues one must confront were all confronted historically. My analysis will draw primarily on the foundational thinking of the Greeks and, to a lesser extent, on one of the inventors[2] of analytic geometry, namely, Rene Descartes.

It is not possible to fully understand how *mathematical* concepts tie to reality if one does not understand how *concepts in general* apply to reality. Because Ayn Rand's theory of concepts has provided that understanding, her theory of concepts provides a fundamental underpinning of this study. My treatment will not contain an exposition of her ideas nor presuppose a prior understanding of them. But I will be applying her framework at every turn, will be looking at mathematical concepts from that perspective. As in Chapter 1, when particularly appropriate, I will call attention to the specific passages of her work that crystallize the underpinnings of my own analysis. It is Ayn Rand's theory, in my view, that enables one to avoid the false alternative between the Platonic view of a separate mathematical universe and the more modern view that mathematical concepts are arbitrary conventions requiring no existential referent.

Some puzzles go unnoticed at all levels of education, from grade school through

Some puzzles go unnoticed at all levels of education, from grade school through graduate school. In our quest, we will encounter, and address a number of them. For example:

● Why is a ratio different than a fraction? Or is it?

● Do we need a mathematical infinity of numbers because the world is infinite or because it is finite and measurable?
● What does the unit '1' on the real line represent? This one may be the most puzzling, because it sounds like a silly question. So the first puzzle really is: Why isn't it a silly question?

A final question would require a knowledge of the history of mathematics that is not part of the standard curriculum:
● If Descartes was right about Cartesian coordinates, was it for the right reason? Or for the wrong reason?

I've worded these questions as teasers and, in what follows, I do not ask these questions in quite this way. They do not arise in conventional treatments of real numbers. But they arise naturally when one begins with the phenomena that real numbers are needed to represent, namely the measurement of magnitudes.

It is not my intention in this chapter to discuss standard approaches (such as the mathematical constructions I once learned as a mathematics student) to understanding the real line. My discussion of Dedekind and Cantor regarding the real numbers is deferred to Chapter 4. My task in this chapter is a more positive one, namely to understand the real line: To present a reality-based approach to understanding the real numbers, specifically as they apply to magnitudes. I will address the questions along the way that have aroused my own interest and that I consider essential to a full understanding of elementary arithmetic.

I will begin, as I say, not with real numbers, but with what they measure. I will begin with magnitudes, such as the length of long objects and the speed of moving bodies. I will look at magnitudes the way one had to when measurement was still being invented. And I will take a geometric perspective in the spirit of Euclid's *Elements* because such a perspective is a way of turning the spot light away from the quantitative abstraction, as such, and toward the genesis or content of such abstractions. As I see it, the geometric perspective, broadly conceived, is a way of focusing on the content of mathematical abstractions, on what is being measured. The geometric perspective helps distinguish what is

being measured from the means by which one measures.

Geometry, Measurement, and Magnitude

To properly understand the measurement of magnitudes one must first look at them as the ancient Greeks did. One must look at them geometrically.[3]

The geometric perspective directs ones focus toward the object of measurement, as opposed to the numerical results of such measurement. One thinks of the objects as having an independent existence, as having preexisting relationships to other objects, and as being measurable.

When I look at a distance as being five miles I am not looking at it geometrically; I am looking at it numerically. When I perform an algebraic calculation involving numerical unknowns, I am also not thinking geometrically. But if I compare two pencils with regard to length and say that this pencil is longer than that or if I say that this pencil is three times as long as that, I am looking at it geometrically. My focus is on quantitative relationships between the pencils. This remains true whether or not I draw lines to represent the pencils.

In general, looking at an object quantitatively, but without regard to a specific, given unit of measurement is looking at it geometrically. But this is not a criterion and it can be a matter of emphasis. For example, if I consider the relationship of a choice of unit to the numerical results of that choice, I need to retain my focus on the object of measurement. In such a case I would be thinking geometrically. For example, when one converts feet to yards, one presupposes, and focuses on the fact, that one and the same object is being measured in two different ways.

As in Chapter 1, I will examine the underpinnings of measurement, this time as it relates to magnitudes. I will examine the quantitative relationships that make measurement possible, possible in the full sense of “measurement” articulated by Ayn Rand. As a reminder:

“Measurement is the identification of a relationship–a quantitative relationship established by means of a standard that serves as a unit.”[4]

Chapter 1 focused on the preconditions of geometric measurement; this one will

focus on the preconditions of the measurement of magnitudes. My analysis will look at measurement in the general way that Euclid (implicitly) did, as determining or specifying quantitative *relationships* in an abstract setting that apply to an open-ended range of concretes. In chapter 1, I characterized this as abstract measurement.

But a quantitative relationship is a relationship between two quantities, two quantities of a particular characteristic. When one isolates a similarity, one is recognizing a quantitative dimension along which the concretes vary. In the similarity, one recognizes a characteristic that the concretes share, though in possibly different degree. The specific degree of the common dimension is its quantity.

Quantity is an aspect of the identity of the characteristic, an aspect that we identify either by direct perception or by relating it quantitatively to something that we do perceive directly. As such, quantity is something that exists in the world. Quantity is not the number or specifications that we attach; rather, it is that to which we attach the number or specifications.

The geometric perspective taken by Euclid is a focus on quantities as objects of measurement, as the objects of study and as relatable quantitatively, but without regard to a specification of a particular perceptual standard.

A magnitude is a type of quantity. One needs to distinguish magnitudes from the broader category of continuous quantity, on the one side, and from multiplicities of discrete entities on the other. Discrete entities, taken together, comprise a collection. A count of a collection of entities always yields a whole number. By contrast, continuous quantities are not collections, but admit of gradations. A length is a continuous quantity; a collection of books is a multiplicity. A magnitude, in general, is a continuous quantity.

Magnitudes are distinguished from a variety of different kinds of continuous quantity. On the one side, consider that, for example, length is a magnitude. We can compare lengths in terms of multiplicity. We can say that one length is twice another length or that length C is the sum of length A and length B.

In contrast, hardness is a continuous quantity, but it isn’t a magnitude, at least as it was measured in the past.[5] Traditionally, one measured hardness of minerals and gem stones along a comparative scale called Moh's Scale of Hardness. One

said that a diamond had a hardness of 10 while quartz had a hardness of 7 because a diamond can scratch a quartz crystal but a quartz crystal cannot scratch a diamond. Between quartz and diamond in hardness are Topaz and Corundum. The differences were ordinal. There was a way to determine relative hardness, but we could not say, on such a basis, that a diamond is twice as hard as a quartz crystal.

Lengths, areas, weight, acceleration (in a particular direction), force (in regards to the strength of the force), density, and water pressure are all magnitudes. The pitch of a sound is also a magnitude because one can relate pitch to the frequency of a vibration. But we do not perceive it that way: when one vibration is twice the frequency of another we perceive the difference as a musical interval, specifically as an octave. Finding frequency as a unit of measure was a scientific discovery.

Ordinal measurement of physical quantities is unusual, but there are many, more common kinds of continuous quantities that are not magnitudes. For example, force, considered as acting in one of a range of directions, is not a magnitude. It certainly *involves* a magnitude, namely the strength of the force, but to fully specify, to fully measure, a force, one need also identify the direction in which it acts. Many physical quantities are of this type, including velocity (as opposed to speed) and acceleration (conceived as a rate of change of velocity).

Direction falls into yet another important category of continuous quantity. Direction, as such, is not a magnitude; one cannot speak of multiples of north and one cannot add north and east. Northeast is not the sum of north and east, for example. Rather, it is a third direction *between* north and east.

But a *difference* in direction, if measured as a rotation, *is* a magnitude. This is clearest when one's study is confined to a plane. As we saw in Chapter 1, a difference in direction is an angle, an amount of turning or rotation. An amount of rotation is a magnitude, a magnitude that Euclid characterized as an angle and measured in multiples or fractions of a right angle.

In three dimensions, an angle, an amount of turning, can occur in different directions. For example, one can rotate 90 degrees from north to west or one can rotate 90 degrees from the vertical to the horizontal. Rotating involves both the amount of turning (which is a magnitude) and the axis of the rotation (which is not a magnitude). A rotation in three dimensions is not a magnitude. But even here, a difference in direction, the angle,

considered without regard for the axis of rotation (an omitted measurement)[6] *is* a magnitude.

There are many kinds of quantities that are not themselves magnitudes but for which differences between concrete instances are magnitudes. Direction is a striking example of this phenomenon, but is only one of many important examples. Another important example is position, position along a linear direction. Although length, distance, and displacement (along a linear dimension) are magnitudes, position, is not. But *relative* position, measurable by the length of a tape measure, say, stretching from one object to the second and without regard for direction, *is* a magnitude. Displacement, a movement of something from one position to another, involves that same magnitude. Time intervals, although harder to measure, are a similar example. In physics, measures of potential energy and electric potential require selecting a zero-point. What one measures, as *magnitudes*, are *differences* in potential energy or *differences* in electric potential.

Examples such as direction, vectors, and position follow a general pattern. Such quantities are not magnitudes but are measured by magnitudes. In general, such measurements require choice of an axis of measurement and/or a zero point. In the case of vectors, one chooses an entire coordinate system. Choosing a coordinate system for such quantities is exactly equivalent to choosing a standard (such as a meter) to measure length. The numerical values of the numerical measurements (or coordinates) depend upon the choice of coordinate system and are only meaningful in relation to the chosen coordinate system.

The use of magnitudes to measure differences also introduces a refinement into one's conception of magnitude. To measure position, one very often selects a zero-point, a sort of "home" position, like a terminal point on a railroad line. One measures position as the relative position from the zero point. Then one can distinguish the two opposite directions from that zero point, calling one direction positive and the other negative. A difference of position in one direction is positive; in the other, it's negative. As a slightly different example, one might distinguish between counter-clockwise (positive) rotations and clockwise (negative) rotations in the plane.

Notice a difference between these two examples. In the case of position, one's use of numbers to measure linear position involves *two* choices. First, there is

the choice of a home position. Second, there is the choice of a positive direction (which may be to the right) and a negative direction (the opposite direction from the positive direction). But, in the case of rotations, the home position is necessarily: no rotation at all, a zero-rotation. Yet, one still needs to decide whether a clockwise or counter-clockwise rotation (as viewed from a particular vantage point) will be taken as positive. One does not choose the zero point, but one does need to decide which sense of rotation is taken to be positive. Displacement is similar in that regard. A displacement from A to B has the opposite sense as a displacement from B to A. One must choose which sense one takes to positive. And the zero point is the trivial displacement from A to A.

Finally, notice that any type of magnitude can form the basis for a related type of magnitude that involves either positive or negative senses. Consider weight, for example. Ignoring the phenomenon of buoyancy, weight is always positive. But one can compare weights and measure differences in weight. One can compare the weights of Mary and Joe. Joe, let us say, weighs 200 pounds and Mary weighs 120 pounds. Then one can say that Joe weighs 80 pounds *more* than Mary. To paraphrase, the difference in weight between Joe and Mary is 80 pounds, that is, Joe is 80 pounds more than Mary. But one can also turn this around and say Mary is 80 pounds *less* than Joe. One paraphrases this relationship by saying that the difference in weight between Mary and Joe is *minus* 80 pounds, that is Mary's weight minus Joe's weight is minus 80. Viewed in this light, *differences* in weight admit of positive and negative senses. A positive designation means that one needs to *add* something to get from the second to the first whereas a negative designation means that one needs to *subtract*.

It is important to realize that the positive and negative senses of a magnitude are two senses of the *same kind* of magnitude. For example, a difference in 80 pounds can be a positive difference or a negative difference. The first weight can be either 80 pounds greater or 80 pounds less than the second. But either way, the absolute *magnitude*, the amount, of the difference is 80 pounds. There is a great difference between Joe weighing 80 pounds more than Mary and weighing 80 pounds less than Mary, but either way, the *magnitude* of the difference is 80 pounds. And this same point applies to all of my other examples such as distances between objects, time intervals, potential energy, and electric potential. In all such cases, the amount of the difference is a magnitude. Without knowing the sense of the difference, greater or lesser, clockwise or counter-clockwise, one has not fully specified the *difference*. But one *has* specified the degree of the

difference, its *magnitude*.

On the basis of this discussion, a magnitude can be characterized as a continuous quantity that:

1. Admits of comparisons of greater or lesser with other quantities of the same type
2. Can be related to other magnitudes of the same type in terms of multiplicity. That is, one can determine that one magnitude is three times another magnitude of the same type.

In this, one takes for granted that, for example, when one doubles a magnitude one obtains a greater magnitude of the same kind, that magnitudes are divisible, and that when one subdivides a magnitude, the parts are each less than the whole and add up to the whole. I also take the ability to add two magnitudes of a particular type to be characteristic of magnitudes and implicit in the relation of multiplicity. Finally, in my discussion of *magnitude*, I will, much of the time, ignore the distinction between positive and negative.

As the examples indicate, it is sometimes easier and sometimes harder to determine whether a quantifiable characteristic of something is a magnitude and, if it is, to determine how to relate one instance to a multiple of another instance. Difficult or easy, these are discoveries that must be made anew with each new kind of magnitude one encounters. The mathematical theory of continuous magnitudes and of the positive real numbers[7] that measure them becomes applicable once this discovery has been made.

Representing Magnitudes by Line Segments

One of the easiest magnitudes to grasp as a magnitude is the length of an object. Moreover, it is easy to create a stylized, visual representation of length, by a line segment, to help visualize the quantitative relationship one wants to study.

Euclid followed this procedure in Book V of the *Elements*.[8] Euclid's theory of ratio did not limit the type of magnitude to which his theory applied, but required only that the two magnitudes related by a ratio be the same kind of magnitude.[9] And Euclid, explicitly, took ratios between lengths and ratios between areas, starting in Book VI. Yet, throughout Book V, he used line

segments to represent magnitudes, without regard to whether any particular magnitude, thus represented, is a length or an area.

Archimedes went even further, using line segments to represent weight or, implicitly, force/buoyancy in an upward or downward direction. For example, Archimedes's famous law of levers relates weights applied on a balanced lever arm to distances along the lever. To illustrate his derivation, he drew areas to represent weights and line segments to represent lengths. In his derivation, Archimedes draws freely on the methods presented in Euclid's Book V.[10]

Euclid's application of his theory of magnitude to both length and area is critical to his theory. As I will show in Chapter 4, Euclid's theory of proportion requires not only a way to measure the ratio of two areas, but also a way to compare a ratio of lengths to a ratio of areas. Euclid's own applications illustrate that the power of representing magnitudes by line segments consists in the fact that one's discoveries *do* apply to other kinds of magnitudes. These discoveries apply generally to magnitudes because the arguments depend only on characteristics common to all magnitudes.

In what follows, I will use line segments to advance my discussion of the "prearithmetic" of magnitudes. But it will be important to explore, as I proceed, just how the central ideas apply to other kinds of magnitudes.

The PreArithmetic of Magnitudes

Taking the geometric perspective, the first task is to explore the ways that quantitative relationships can be grasped without first choosing a standard of measurement. Such relationships do not, could not, be *created* by the numbers we attach to them. It is quite the reverse. One can apply numbers to magnitudes because the relationships they bring to light already exist. And one's ability to relate a magnitude to a chosen universal unit already presupposes a prior, more general, ability to compare magnitudes. One's ability to identify a length as being twice the length of a foot does not depend on having chosen that foot as one's unit of length.

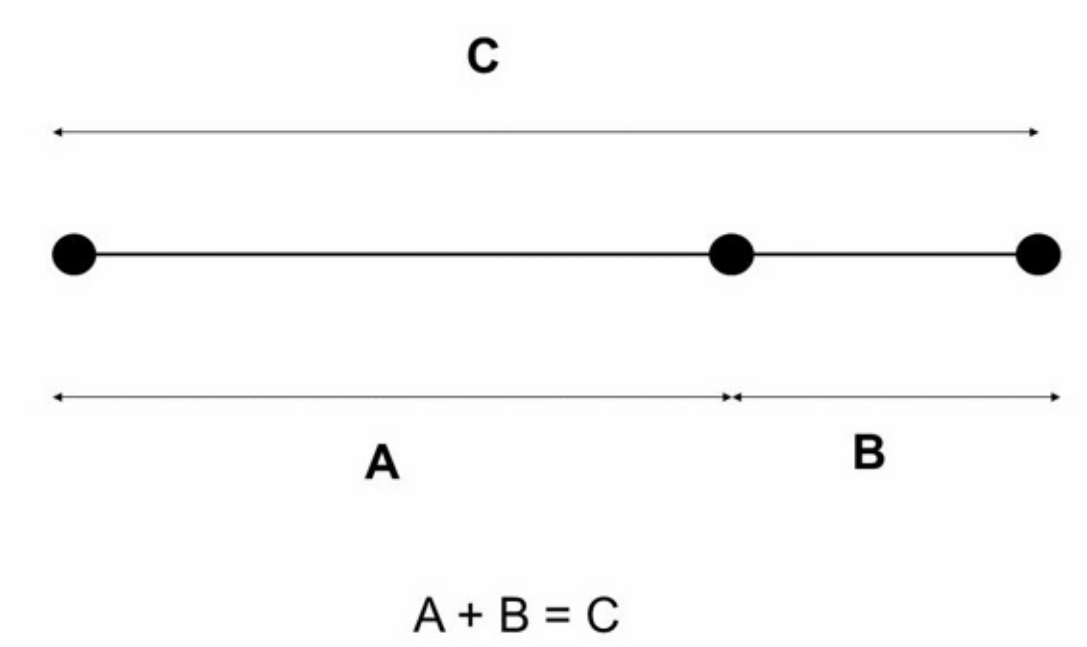
Even so, it may seem unnatural to think about the relationships between quantities without attaching numbers to the quantities being related. But one needs to remember that relationships between quantities are real. These

relationships exist before one measures them and are *discovered* during the process of measuring.

In this regard, remembering one's study of plane geometry may help to establish the right frame of mind. That entire subject is usually taught with nary a number in sight, yet the subject abounds with bisections, equalities, and ratios. Remember, as well, that the Pythagorean Theorem, as conceived geometrically, is not about the sums of squares of *numbers*, but of the sums of actual *squares* erected on the sides of a right triangle.[11]

I will apply Euclid's method, discussed in Chapter 1, to the study of magnitudes, as well. I will focus on quantitative relationships. I will not offer straight edge and compass constructions, as Euclid does. But I will retain the essence of this approach by elucidating, as appropriate, how one might determine various quantitative relationships. Nonetheless, in cases for which the ability to perform a particular measurement is sufficiently well understood, I will take these measurements for granted.

Like arithmetic, what I call the prearithmetic of magnitudes begins with addition.
Magnitudes can be added. Sometimes, as with line segments, one can say this quite literally: Lay them end to end and you have a longer line segment, the sum of the two. In the diagram below, line segments A and B are added to form a line segment C. Please keep in mind that the letters "A", "B", and "C" do *not* stand for numbers in this picture; they name geometric objects, line segments taken as open-ended abstractions:



To emphasize, one constructs this sum geometrically without ever saying anything like “This length is 4 inches, that one is 5, therefore the total is 9 inches.” Inches or centimeters do not enter into this in any way. Moreover, the geometric construction I have outlined is simply an abstract way of specifying what one would do in each concrete instance. Namely, in the case, say, of pencils, one lays one pencil end to end with the other to find, physically, the sum of their lengths. As I have stated, looking at quantitative relationships in this way is the essence of the geometric perspective.

But such a procedure is not always available. For example, how does one add two frequencies? I will answer this question in a moment, but clearly a more abstract perspective on this is needed, even when I’m not talking about numbers.

So, to continue, when I talk about adding magnitudes, what I’m really defining is a relationship. I’m defining a relationship between pairs of lengths A and B versus other lengths equal to the physical sum C of A and B. In the case of line segments, I lined up two lengths end to end to get their sum, the length C. But any other magnitude that would match this sum, any other magnitude of length C, has a length equal to the sum of lengths A and B. The sum of lengths A and B is not specifically the length of one particular physical object. Rather, it is that

characteristic, the specific length, that all lengths of that particular combined length have in common.

By the same token, any length greater than C is greater than the sum of A and B and any length less than C is less than the sum of A and B.

Moreover, in speaking of the sum of A and B, I do not claim that there actually exists, at this moment, a particular physical object of length C, the length that would result by laying A and B end-to-end. Nor have I, in fact, created such a length. Without actually carrying it out, I have specified a recipe, an abstract measurement, for doing so. And the meaning of the result, the conceptual units of my prescription as applied to particular lengths A and B, consists in all the magnitudes that ever existed, that ever will exist or that might exist with a length of C. What is important is that I have *specified*, by reference to a chain of abstract measurements and subject to contextual precision requirements, those magnitudes of length C equal to the sum of the magnitudes A and B.

The actual way that one adds magnitudes depends upon the particular magnitude in question. The appropriate method must be discovered in each case. Weights, for example, are added by bringing them together in some way, perhaps attaching them, placing them on the same balance scale, or causing the weights to join forces in some other way.

A different approach is required to add speeds. In this example, speed must be taken to be constant during a particular interval of time. Speed involves a distance traveled within a certain interval of time. If one travels twice the distance in that time, one obtains twice the speed. More generally, to add speeds, choose any convenient time interval. Having chosen the interval, one simply relies on the fact that addition of distance has already been defined. So the sum of the two speeds is simply the speed for which the distance traveled during the time interval is the sum of the two distances for the speeds being added.

Notice that, unlike the case of length, I have not prescribed laying the speeds end to end; I have simply appealed to that process, for lengths, *to specify* the meaning of a sum of two speeds, thereby prescribing the *physical relationship* between the summands and the sum. It is that specification, not the means by which it is specified, that matters. The entire *specification* is an abstract measurement in the sense I defined in Chapter 1.

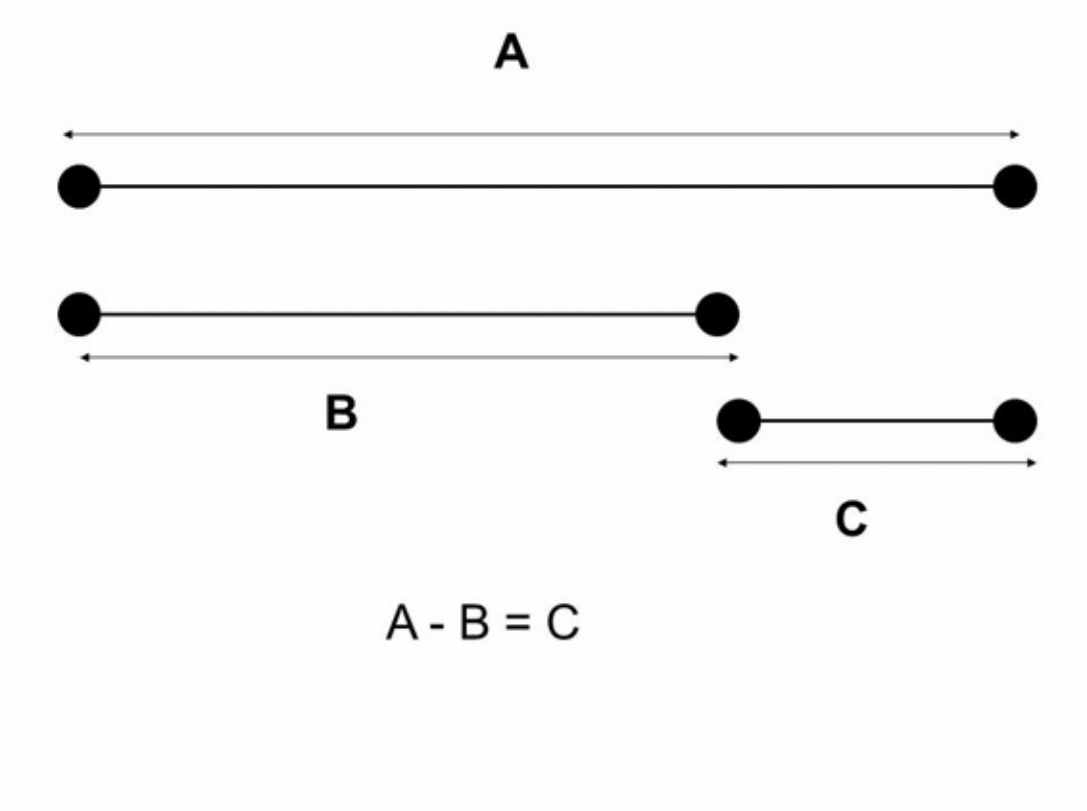
Pitch is a similar case. A pitch is a manifestation of a certain multitude of

vibrations during a specified period of time. The sum of two pitches is that pitch for which the multitude of vibrations during a specified time interval is the sum of the multitudes for the two summands. (The sum of two multitudes is simply the combination of the two multitudes considered as a single multitude.) In this case, to understand the addition of multitudes, both what it entails physically and what perspective it entails, it is enough to specify the sum of the two pitches; such a specification counts as an abstract measurement. However, this specification does *not* prescribe a way *to construct* such a pitch, even in pattern. For such a prescription would have to involve something like the following (changing the example slightly): If a string has a particular thickness, tension, and length, these three factors determine the pitch of its vibration when it is plucked. If one keeps the thickness and tension the same, but cuts its length in half, one doubles the frequency of the pitch and hears a tone one octave higher. (On a violin or a guitar, one does this routinely by pressing the midpoint of the string with one's finger.) But this is a physical *implementation* of a *prescribed* quantitative relationship, a relationship that has been *defined* without having specified the *means* of bringing it about.

In general, many physical characteristics are measured by their effects on motion. Such effects are generally quantifiable with respect to displacement, time, and force. One measures physical characteristics such as electric charge indirectly by measuring their more-directly measurable effects. Even force can often be quantified by the effect, in regards to its motion, on a particular kind of object. In such cases, for those physical quantities that are actually magnitudes, one identifies a method of addition with reference to the methods for adding distance, time, and force. My discussion of adding speeds followed this pattern.

Using line segments to represent magnitude generally is valid because, and insofar as, the relationship established for the lengths of line segments correspond to similar relationships for other kinds of magnitudes.

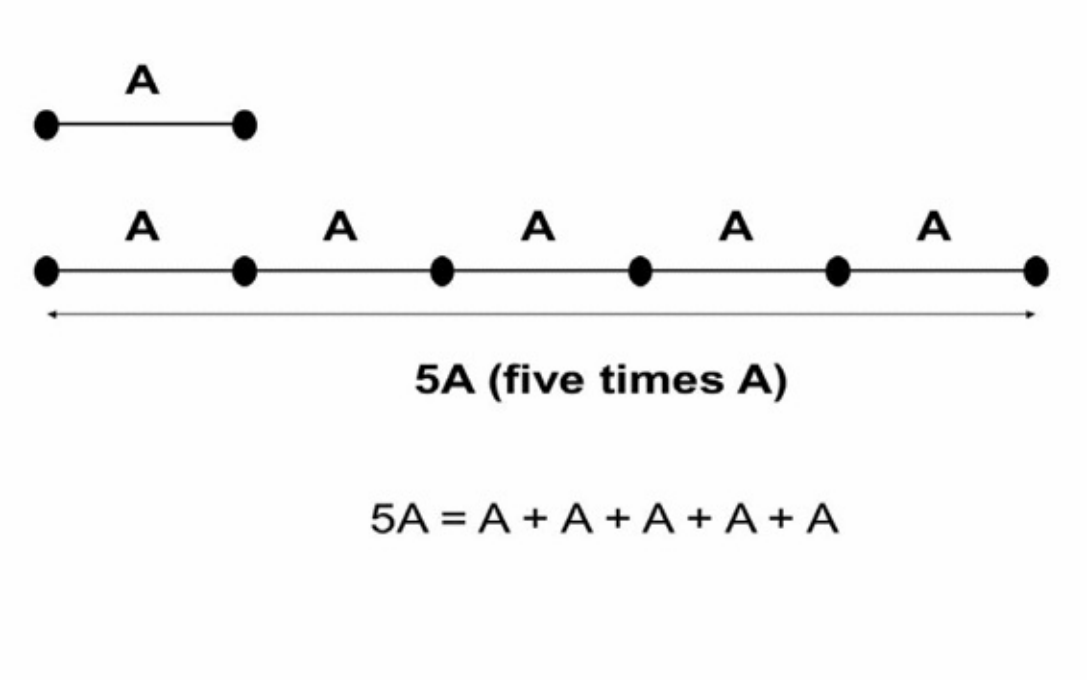
Similarly, one can subtract two magnitudes. Lay two line segments next to each other, matching the endpoints on one end of the line segments. Looking at the other ends, the segment between the endpoint of the shorter and the endpoint of the larger is the difference between them. Again, one compares pencils, finds the difference in their lengths, in precisely this fashion. In the diagram below, the difference between line segment A and line segment B is the line segment C:



For other types of magnitudes, once one has figured out a method to add magnitudes of that type, one simply follows the inverse process for that same method, as I have indicated here for the subtraction of lengths.

Next, one can multiply a magnitude by a number. Here, I warn the reader, in advance, that this is *not* the same as multiplying a magnitude by a magnitude. Rather, as we once learned in grade school, multiplication by a number can be a short-hand for repeated addition. To multiply a magnitude by five, add together a total of five repetitions of the magnitude. Numbers are being used, here, to count; the numbers count repetitions.

Symbolically, one might write $5 \times A = A + A + A + A + A$. But one thinks of this operation as being carried out physically. One thinks of a second physical quantity related to the first physical quantity in the indicated fashion. In the case of line segments, one can count repetitions of a length and one can apply arithmetic to the numbers used to count these repetitions. But one cannot use a number to represent a *length* and add *those* numbers until one has chosen a standard such as feet or meters. There are no numbers to add without a standard. So in the next diagram, 5 times the line segment A comprises 5 line segments of length A laid end to end:



To recap, the number “5” has entered the discussion. What does “5” represent? Well, it has nothing to do with whether we measure the length A in feet or meters or whether we ever *measure* the length A at all. Rather, “5” represents nothing more nor less than the number of repetitions, the number of times that the length A is laid end to end with itself; the number of occurrences of “A” in the formula above. The process described is totally independent of any standard of length that one might select to measure the line segment A.

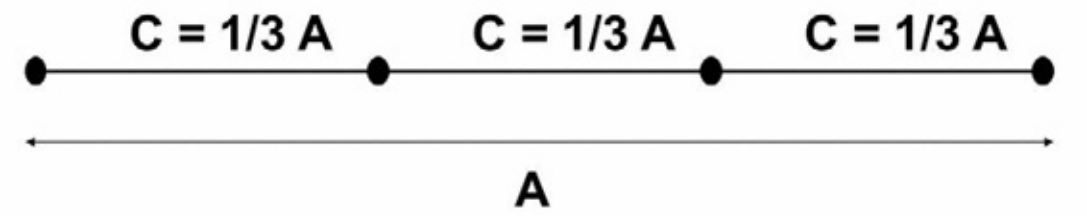
To clear up a possible confusion, there is one limited sense in which “A” functions as a standard in this example. Namely, when the iterations of its length are counted, it is treated as a standard in exactly the way that counting always treats the units that are being counted. To wit, it is one of the units being counted. “A” happens to be the particular length that is being laid end to end. But this does not make it a standard of length.

One should also note, here, that my method of multiplying a magnitude by a number presupposes that one knows how to add two concrete instances of that magnitude. And this is the proper hierarchy: to define multiplication in terms of addition. So, when I characterized magnitudes as being relatable to each other in terms of multiplicity, this characterization presumed the ability to add magnitudes. In sum, the ability to add magnitudes, as I mentioned in that discussion, should be taken as implicit in my characterization.

On the other hand, the ability to add quantities of a particular type does not make them magnitudes. For there are obvious counterexamples. For example, one can add velocities and one can add forces, something that physicists properly do routinely. But velocity is characterized by both a magnitude (its speed) and a direction. If two velocities in different directions are added, the speed of the sum will be less than the sum of the speeds of the summands.

This treatment does indicate that once one has defined addition for a particular kind of magnitude, one has a general process for defining subtraction of magnitudes of that kind, multiplication of them by numbers, and, as we shall see shortly, division by numbers, as well. Once the additive relation has been identified for a particular kind of magnitude, the rest of my discussion applies to that kind of magnitude, as well.

Just as subtraction relates to addition, division relates to multiplication. Thus, dividing a magnitude into three equal smaller magnitudes means finding a magnitude that, when multiplied by three, yields the original magnitude. In the following figure, C is one third of A precisely because 3 times C equals A:



$$A = C + C + C = 3C$$

$$C = 1/3\ A$$

Now the ancient Greeks had a geometric construction, using a straight edge and

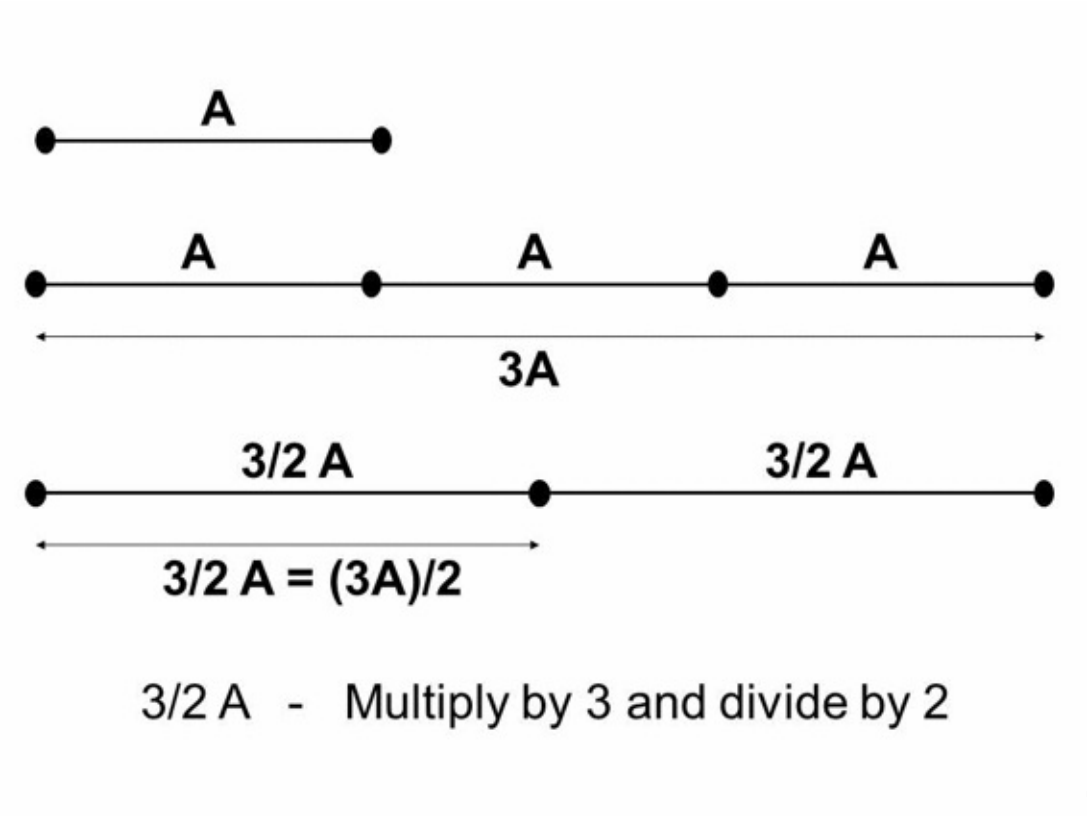
compass, to divide a line into any prescribed number of equal segments. But they tried unsuccessfully to find such a construction to subdivide angles. They could bisect them, but to trisect them (divide an angle into three equal angles) required mechanical contrivances that they considered unsatisfactory.[12]

In general, one can specify mathematically what it means to subdivide a magnitude. But the actual physical process to carry out this division is a separate discovery for each new magnitude that one discovers. As in the example of sound pitch (vibration frequency), this subdivision can be highly non-trivial. This is especially true when, as in that case, even the addition of magnitudes (frequencies of vibration) is non-trivial, because the identification of the particular kind of magnitude that underlies our observations requires a scientific discovery. Keep in mind, as well, that there can be physical limits, for any particular type of quantity to which a division can be carried out. As two examples, as far as we know, Planck's constant is the absolute minimum of action that is physically possible and 1/3 of the charge of an electron is the smallest unit of charge that is known to exist. So, for example, one can specify how one fourth of the charge of an electron would relate to the charge of an electron. We would recognize it if we found it. However, to our knowledge, there is no such subdivision.

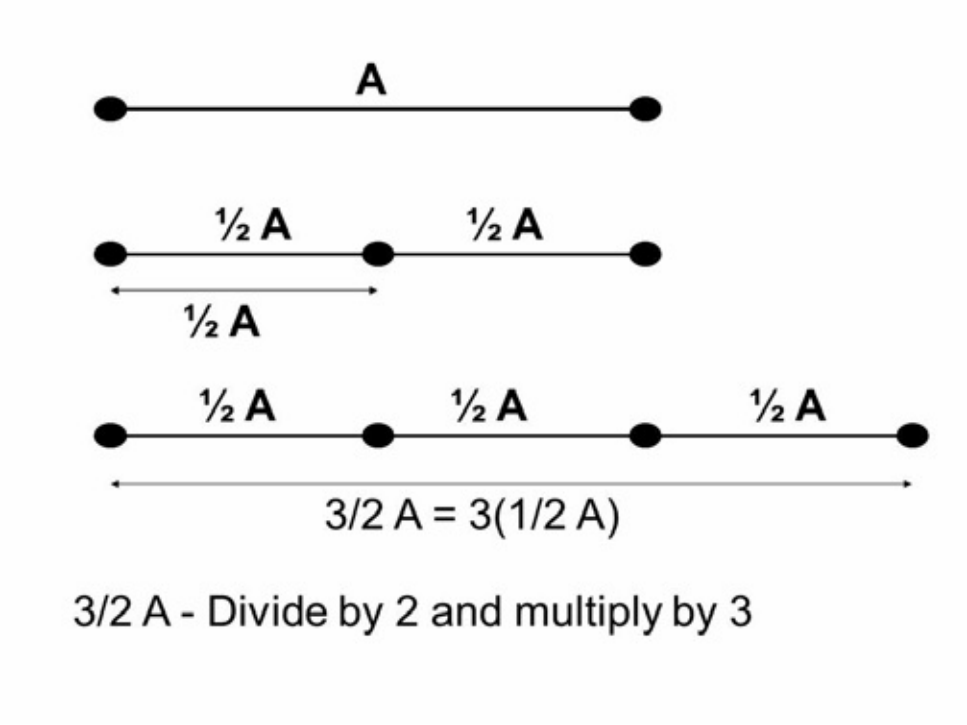
Indeed, there is a significant difference between the way that one subdivides a length and the way that one subdivides most other magnitudes. For example, one does not subdivide a pitch into two pitches nor, for that matter, does one physically add two pitches together to create a new pitch. One only has different pitches from two different objects vibrating at different speeds. Nonetheless, what one *can* do is to relate two different pitches. One can say, for example, that the pitch one octave above middle C has twice the frequency of middle C. Or one can establish that the sum of the frequencies of two different pitches is equal to the frequency of a third, a judgment that is independent of any standard of measurement that one might have selected. These relationships between pitches are relationships of *magnitudes* just as the relationships of line segments are relationships between magnitudes and the relationships are the same. What differs in the two cases is the *means* of establishing those relationships and the specific *form* that they take.

I have now discussed multiplication of magnitudes by whole numbers and division of magnitudes by whole numbers. Putting these together, one can interpret multiplication by a fraction N/M to be the magnitude obtained by first multiplying it by N and then dividing the result into M equal pieces. Each one of

multiplying it by N and then dividing the result into M equal pieces. Each one of the resulting pieces has the desired magnitude. The result is the same as first dividing the original result into M pieces and then picking one of these pieces to multiply by N. This first method is illustrated below for N = 3 and M = 2.



The second method:



This is the same process one follows in arithmetic: To multiply by a fraction, multiply by the numerator and divide by the denominator in either order.

The process of multiplying by an irrational number works for magnitudes the way it works for numbers. Recall that an irrational number, such as the square root of two, is a number that cannot be expressed as a fraction, as a ratio of whole numbers. In practical life one deals with irrational numbers by finding a suitably precise approximation of the irrational number by a rational number

That approach applies here, as well. In any concrete context, there is a limit to the precision that is needed and to the precision that can be achieved. Subject to such limits, one finds a suitable rational approximation and then multiplies the magnitude by that rational number in the way that I have described.[13] Within the specified precision context, the resulting magnitude bears the required relationship to the multiplied magnitude.

The full justification of this process requires that it is always possible, no matter how demanding the required precision, to find a qualifying rational approximation. That such an approximation can always be found is a consequence of the Axiom of Archimedes, discussed in the next section.

Multiplying by irrational numbers, in this way, assumes that the irrational

number has already been specified. In general, one specifies irrational numbers in relation to other, already specified numbers. So, for example, the square root of two, an irrational number, is specified in relation to the number 2: It is that number whose square is 2. The general problem of specifying irrational numbers is the focus of Chapter 4.

To recap: Without ever specifying a unit of measurement, one can relate a magnitude as being the sum or difference of two other magnitudes. As a consequence, one can also multiply a magnitude or divide a magnitude by a number. In all four cases, addition, subtraction, multiplication, or division, the result is another magnitude of the same kind.

This provides a first glimpse of the relationship of geometry to measurement. We are all accustomed to using yardsticks and tape measures. To measure something one simply stretches the measuring tape along the length that one is measuring. But where did the measuring tape come from? How is it that one is able to lay off a marking every foot of its length and then subdivide one of these feet into inches, half inches, and quarter inches, and then repeat these subdivisions between every other foot-marking of the tape measure? Because: We already know how to add lengths together and we already know how to subdivide them in just the ways that we have described.

To conclude: The geometric perspective precedes measurement. To measure is to measure *something*. Before one can measure magnitudes one needs to know how to add physical magnitudes and how to subdivide them.

The Axiom of Archimedes

In Aristotle's *Physics*, as part of his argument against the existence of actual infinite magnitudes, one finds:

"…for every finite magnitude is exhausted by means of any determinate quantity however small."[14]
Today, Aristotle's statement is known as the Axiom of

Archimedes,[15] after the greatest mathematician of antiquity, considered one of the very greatest of all time. Keep in mind, though, that Aristotle preceded both Euclid and Archimedes, though he didn't precede Eudoxus. Whatever else may

be true, Archimedes did not originate the axiom of Archimedes and Aristotle grasped its import.

To paraphrase, given any two magnitudes of the same kind, one can obtain a magnitude that exceeds the larger by taking a sufficiently high multiple of the smaller magnitude. If A is the smaller magnitude and B is the larger, there is some whole number N such that N times A exceeds B.

The very statement of this axiom requires the geometric perspective on magnitude that I have been developing. Even in his brief statement one notices that Aristotle takes that perspective entirely for granted.

I will need another form of the same axiom. Applying the axiom to A and B, suppose that one has found a positive integer N such that

$N \times A > B$

Subdividing the segment $N \times A$ into N segments yields N segments of length A. (Indeed, $N \times A$ is already subdivided by virtue of consisting of N segments of length A laid end to end.) Subdividing B into N segments yields a magnitude that one can write as $1/N \times B$ or, for short, B/N.

Now the subdivision of a larger magnitude must exceed the same subdivision of a smaller magnitude. So evidently,
$A > B/N$

As a reminder, this statement means, and only means, that when or if B is subdivided into N equal pieces, each of these pieces is less than A. I emphasize, again, that this is *not* a statement about numbers, but about magnitudes.

Now it may happen that subdividing B into N equal pieces is not actually possible. So a precise statement requires a formulation to the effect that if B were subdivided into N equal pieces, each of these pieces is would be less than A.

Thus, I state a second form of the Axiom of Archimedes: Given any two magnitudes, a sufficiently fine subdivision of the larger into a finite number of equal pieces, if carried out, would consist of pieces smaller than the smaller magnitude. If A is the smaller magnitude and B is the larger, there is some whole number N such that B/N is less than A.[16]

Why care?

Because the actual meaning of the Axiom of Archimedes is that all magnitudes are measurable. As will become evident, to explain this will require both forms of the axiom.

It is said that all journeys begin with a single step! If one takes inspiration from this, it is from the unspoken premise that the journey can be completed with a finite number of steps, that each step takes us one step closer to one's destination.

Consider two distances. Choose the smallest and think of it as a standard. For purposes of discussion, say the smallest is a mile. The first form of the Axiom of Archimedes states that some multiple of the smaller magnitude, which one might choose as the standard, will exceed the larger one. For some large number N, one can say that the larger distance is less than N miles.

Were this not the case, if the distance could not be bounded by any specific number of miles, the distance would be infinite. One would not be able to relate it to the chosen standard; one would not be able to measure it.[17]

From this one sees that measurement of magnitudes, one's ability to measure all magnitudes of a particular type, however large, presupposes the Axiom of Archimedes.

Keep in mind that, for two magnitudes to be magnitudes of the same kind they must be commensurable . In Ayn Rand's terms, they are same characteristic, "but in different measure or degree."[18] Two lengths differ in degree; a weight and a length differ in kind. One can measure one length against another because they are commensurable. One cannot measure a length against a weight because they are not commensurable. We see, in all of our observations of the world, a relationship between commensurability and measurability. And contrary to all of this experience, the failure of the Axiom of Archimedes would imply that magnitudes can be commensurable[19] without being measurable.

Mathematics, however, cannot offer a proof of the Axiom of Archimedes.[20] Knowledge of that axiom is something that one brings to mathematics. One's universal experience is of finite quantities and one sees this axiom as naming a basic fact, inherent in the nature of the world. In essence, it is an aspect of the law of identity. Things are finite because they are limited; because their characteristics are specific; because they have a specific nature.[21]

How does one appeal to this axiom when one measures?

Assume that one has chosen a unit of distance. Suppose D is a distance and U is the chosen unit. Then the Axiom of Archimedes states that there exists some whole number M such that

$D < M \times U$

This is another way of saying that the distance D, expressed numerically in terms of our chosen unit, is less than M. Next, there must be some lowest whole number N such that

$D < N \times U$
Because N is the least such number, it follows that
$(N - 1) \times U <= D$

In other words, D, expressed as a multiple of the chosen unit, is between N – 1 and N. If U is a mile and N is 501 then this says that the distance D is between 500 and 501 miles.

This is already a measurement of D. But one may want to refine that measurement. One subdivides U into tenths of a mile. One may discover, say, that D is between 500.6 miles and 500.7 miles. Depending upon the accuracy required by the context one might continue to subdivide further until one is either unable or unwilling to pursue a more precise determination.[22]

One knows from experience that a decimal expansion in terms of a chosen unit can specify a measure of a magnitude within any required precision. One learns, conversely that there is a limit, in any particular case, to the precision that is physically possible, that beyond a certain number of decimal places any further refinement becomes meaningless, that one reaches, for example, what Corvini calls the distinguishability limit.[23] But within these limits one learns that decimals can approximate any number to any required precision and I've now outlined the process by which this is done. Once a standard has been chosen, any two magnitudes that one is able to distinguish have a different decimal expansion. There is some point at which the numbers in the expansions differ; there is some rational number that lies between them.

The Axiom of Archimedes identifies the principle underlying these calculations. The Axiom of Archimedes asserts, by implication, that any magnitude can be related numerically to any other magnitude of that same type.

To explicitly relate this principle to the prearithmetic of magnitudes, recall my discussion of precision in Chapter 1. Specifically, to say that two magnitudes, of the same type, are equal is to say, in a particular context, that there is no material

difference between them. So I reiterate my claim in the last section that, for any chosen standard of materiality, given two magnitudes X and Y there is a *rational* number A such that there is no material difference between Y and AX. For the sake of the argument assume $X < Y$.

Suppose, then, that X and Y are magnitudes of the same type; suppose that $X < Y$, and suppose a precision context in which any differences less than a particular magnitude $Z > 0$ are immaterial. Since Z is a positive magnitude, the Axiom of Archimedes says that there exists some integer N for which $X/N < Z$. Given the choice of Z, this means that differences of X/N or less are also immaterial.

Now, again by the Axiom of Archimedes, since $X/N > 0$, there is a smallest number M such that $Y < M \times X/N$, which also implies

$(M - 1) \times X/N <= Y < M \times X/N$
But, since
$(M - 1) \times X/N = ((M-1)/N)X$ and $M \times X/N = (M/N)X$
It follows that
$((M-1)/N)X <= Y < (M/N)X$

Since Y is between two magnitudes that differ by the immaterial amount X/N, Y is equal, within the assumed standard of precision, to either term or, indeed, to any magnitude in between. Setting $A = M/N$, there is, in this context, no material difference between Y and AX. In this precision context, then, $Y = AX$. One says that the ratio Y to X is A, and also uses the notation $Y/X = A$ as expressing the same relationship in a different form. One thinks of Y/X as designating the ratio of Y to X, as representing the number, A, for which $Y = AX$. In showing how to find a suitable value A once a context is given, in thus providing a recipe to specify A, within materiality for *any* context, one has specified it for *all* contexts.

But now consider the question of the uniqueness of the rational number A, suggested by the expression $Y/X = A$. Once a precision context has been specified, can one say that, in any sense, that there is a *unique* rational number A for which $Y = AX$? Is there, as my notation suggests, a unique number A such that $Y/X = A$? If the question is asked within a specified precision context, then the answer is no, indeed clearly no: The differences among satisfactory values are immaterial and one can say that A is determined up to materiality, but A is

not unique. Indeed, in terms of my construction, any rational number A = p where p is between ((M1)/N) and (M/N), would also satisfy Y = AX, would provide a value of AX that could not be distinguished from Y. These differences in values of A may be immaterial, as measurements of Y with respect to X, but they are, nonetheless, different as numbers, as specifications of relationships. So cases such as this, *when the required precision level has already been chosen*, are not an appropriate setting for this question of numerical uniqueness. In such cases, the proper concept is not numerical uniqueness, but uniqueness up to materiality.

The appropriate mathematical context for *numerical* uniqueness arises only when one looks for a numerical solution that is independent of precision context. That is to say, it arises when one *specifies* a relationship independently of precision context.

The question arises, for example, when X and Y arise geometrically as, say, the side and diagonal of a square. General geometric relationships, as we discussed in Chapter 1, apply simultaneously to all precision contexts and are indepenent of precision context. So in such cases it makes sense to ask: Is there a unique number A, possibly irrational, that can be said to solve the equation Y/X = A (or Y = AX) regardless of precision context? (In this example, that number would be A = √2.) The geometric problem does not specify a precision context and it requires a solution to the equation that is, accordingly, independent of precision context. In this sort of situation the question of uniqueness does make sense and, in fact, the numerical solution can be shown to be unique.

In Chapter 4, I will explain in just what sense √2, in this example, is a number A that uniquely satisfies Y = AX simultaneously in all precision contexts. For now, notice that, in any particular precision context, successive rational approximations (m/n) do not, ultimately, differ materiallyfrom√2, the ratio between X and Y. Regardless of precision context, there comes a point at which all sufficiently precise rational approximations to √2 solve the equation Y = AX. As we will make clearer and more precise in Chapter 4, this amounts to saying that Y/X = √2 in each precision context, i.e., simultaneously in all precision contexts. Until then, take it that, whenever one first specifies a required precision, then all sufficiently precise rational approximations A to √2 satisfy Y = AX.

But is √2 unique as a numerical solution that applies to all precision contexts?

Continuing the geometric example, assume, to the contrary, two distinct solutions, two numbers $A_1 < A_2$ such that, $Y = A_1X$ and $Y = A_2X$. As I will shortly show, one can apply the Axiom of Archimedes to numbers (as well as to magnitudes). So I conclude that, for any pair of numbers, $A_1 < A_2$, there exists a ratio of integers $p = m/n$ such that $A_1 < p < A_2$. Applying, the Axiom again, one finds yet another rational number q between p and A_2so that $A_1 < p < q < A_2$.

Now assume a sufficiently fine precision context, one able to detect differences of magnitudes >=(q-p)X. In this context, one can detect that (qp)X > 0 and therefore that pX < qX, which, in turn, implies

$A_1X < A_2X$

The assumed two solutions are, therefore, contrary to supposition, distinguishable in a sufficiently fine context. So, to conclude, there is at most one number A that satisfies Y/X = A (or Y = AX) in all precision contexts.

In the forgoing, I have shown existence of a solution relative to any prescribed precision context and, secondly, I have now shown uniqueness *across precision contexts* for relationships that arise in a general geometric context. But I have not yet made the general argument to show the *existence* of a number (in general, an irrational number) that si*multaneously* solves an equation Y = AX for all precision contexts.

Now one should expect there to be such a number because any number A that satisfies Y = AX in one precision context will also satisfy it in all less demanding precision contexts. Or, to put it another way, any *interval* of satisfactory solutions for one precision context will also satisfy all less demanding precision contexts. So such intervals are *nested* in the sense that the interval corresponding to a refinement of one's precision standard is contained in the precision interval for the previous precision standard.

However, I am, so far, missing a step showing, for example, that there is, therefore, a non-empty intersection of these precision intervals, that there is at least one number contained in all of these intervals, that will always solve the equation regardless of precision context. I need to show, in general, that any nested sequence of intervals contains at least one real number that is simultaneously contained in each interval. But to proceed further, at this point, would require a far more extensive discussion of irrational numbers.

Accordingly, I defer this discussion to Chapter 4.

In general, I have been looking at magnitudes geometrically and my discussion is about measuring magnitudes. I have not been focusing on the *numbers*, on identifying the means of measurement, on presenting the real numbers as an exhaustive system of measurements. My analysis of irrational numbers and the warrant for considering them numbers, again, is deferred until Chapter 4.

But we needed to apply the Axiom of Archimedes to numbers, just now, in order to maintain that the equation Y/X = A has at most one numerical solution that is independent of precision context. And, in any case, it is of interest to notice what the Axiom of Archimedes says about numbers.

So, in the remaining part of this section, I take for granted that one knows generally about rational and irrational numbers. As a reminder, there are two kinds of numbers. Some numbers, the rational numbers, can be expressed as the ratio of two integers. 6 is a rational number; so is 6/13. But not all numbers are rational; not all numbers can be expressed in this way. For example, the square root of 2, as the classical Greeks had already discovered, is not (in modern terminology) a rational number. Any real number that is not a rational number is an irrational number.

One normally uses decimal expansions to approximate real numbers, to distinguish one real number from a different real number. The basic fact underlying the use of these decimal expansions is that any two distinct real numbers can be distinguished by a rational number lying between them. If A and B are real numbers, then there is a rational number R such that $A < R < B$.

One's warrant for saying this is the Axiom of Archimedes. In its second formulation, as applied to numbers, the Axiom of Archimedes states that, for any number $X > 0$, there is a fraction, 1/N, such that

$1/N < X$

The Axiom of Archimedes, in essence, applies to numbers because it applies to magnitudes and numbers are the measure of magnitudes as they relate to a standard. To choose a standard is, by implication, to assign a number to each magnitude. Thus, to find a multiple of one number to exceed another number *is* to find the multiple of the magnitude corresponding to the first number that exceeds the magnitude corresponding to the second number.

The need to apply the Axiom of Archimedes to numbers reflects a broader point: The need to distinguish numbers reflects the need to conceptually distinguish (and also relate) magnitudes, because that is what numbers do.

Suppose the Axiom of Archimedes did not apply to real numbers. Then one could find a real number X smaller than any conceivable rational number. Then, since X/2 is even smaller than X, the entire interval between X/2 and X, including the endpoints, would consist entirely of irrational numbers, all greater than zero, but less than any assignable rational number.

An entire interval consisting entirely of irrational numbers! How would one ever approximate an irrational number lying within this interval? How could we ever distinguish one such number from another or apply them to the measurement of magnitude? How would one come up with a decimal expansion converging to one of these irrational numbers, as opposed to a different one? Well, one couldn't. One's confidence in decimal approximation is based entirely on the implicit acceptance of the Axiom of Archimedes.

Conversely, the Axiom of Archimedes implies that any finite interval of numbers contains at least one rational number. (A *finite* interval is a continuous interval with two distinct end points.)

To see this implication, it is enough to consider positive numbers. So assume an interval with two end points given by positive numbers $X < Y$. The difference, $Y - X$ is a positive quantity. By the axiom of Archimedes there must be a rational number smaller than that positive quantity. So far, then, for some positive integer N

$0 < 1/N < Y - X$
If necessary, replace N by an even larger integer so that it's also true that
$0 < 1/N < X$

Now one knows by the Axiom of Archimedes that some multiple of 1/N, say M/N, exceeds X. Choose M to be the smallest such multiple.

Then one has:
$(M - 1)/N <= X < M/N$

Here, if X is an irrational number, the first inequality is a strict inequality, that is, $(M - 1)/N < X$. Otherwise, if X is a rational number, $(M - 1)/N = X$.

So, to begin with,
$X < M/N$
Now add the following equations:
$(M – 1)/N <= X$
and
$1/N < Y – X$
The result is that
$M/N < Y$

In conclusion, $X < M/N < Y$, which is what one needed to show.
A finite interval consisting entirely of irrational numbers *anywhere* within the field of real numbers would create an absolute, a priori, limit to the precision achievable by a decimal expansion. The Axiom of Archimedes guarantees that this anomaly cannot occur.
The importance of the Axiom of Archimedes, then, is that it is sufficient, and also necessary, to guarantee that every magnitude of a particular kind can be measured by any unit chosen as the standard for that kind of magnitude. Whatever one's choice of unit, any magnitude of that kind can be measured, thereby, to any preassigned degree of precision by a rational number
Now suppose one argues: Well if $X > 0$ is small enough, who cares if there are any rational numbers less than X? But this question actually highlights another absurdity that would result from denying the axiom.
First, if X really represented a number smaller than any rational number, it would be indistinguishable from 0. But suppose one makes X bigger. Suppose, for example, that one multiplies X by a billion. Well, notice that any *irrational* number, multiplied by a rational number, is also an *irrational* number. If the interval between 0 and X consists entirely of irrational numbers, then the interval from 0 to one billion times X also consists entirely of irrational numbers. And this would remain true if I replaced one billion by a billion trillion, or by an even larger integer, no matter what that number might be. No matter how big the multiplier, the result of the multiplication will be an interval consisting entirely of irrational numbers. And, of course, since 1 is a rational number, one could never find a multiple of X that would exceed 1. Or, to put it another way, 1/X if such a calculation made sense, would be higher than any integer one might choose; 1/X would be infinite.
So, once again, the Axiom of Archimedes is a way of saying that all magnitudes of a certain type are comparable. It says that any magnitude of a particular type can be measured in units of any other magnitude of that type.

Multiplication, Units, and Ratios

We are taught to think of numbers as points on a line (the “real number line”), starting from a zero point and extending in two directions, a positive direction and a negative direction. We are taught the operations of addition, subtraction, multiplication, division, and, possibly, the extraction of square roots. The result of all these operations is always another point on that line.

I dispute neither the value nor the validity of this perspective. But it can be misleading, as well, because it suggests that a number is a type of magnitude.

In this regard some of the subtlety has been lost. Lost, but not forgotten. The public school teachers do not see the subtleties and the mathematicians seem to ignore them. But the physicists are forced to deal with them because “dimensional analysis”, keeping track of units and types of magnitudes, is essential to their calculations. Those who use mathematics need to deal with its subtleties whenever those subtleties become relevant to their concerns.

Multiplication of Magnitudes

When one adds quantities, one adds multiples or magnitudes of the same kind. One adds apples to apples; one does not add apples to oranges. But as long as the units are the same, the rules of arithmetic do not depend on what kind of magnitude or multitude one may be adding or subtracting.

The trouble starts with multiplication. One multiplies feet times feet, but the answer is square feet. One multiplies hours by miles per hour and the answer is in miles. One multiplies mass times acceleration and the result is measured as a force.

One does not multiply apples by apples at all. One multiplies the number of apples in a group by the number of groups, but that is a multiplication by a number of repetitions, not by a number of apples. And this is what I was doing in the last section. When I multiplied a magnitude by a number, I got a magnitude of the same type because I was counting repetitions of a magnitude added to itself multiple times. In the case of apples, “six times three apples” is 18 apples. In the case of feet, “six times three feet” is 18 feet.

But “six feet times three feet” has a different meaning and it has a different answer: “six feet times three feet” is 18 *square* feet.

The Greek geometers did not multiply magnitudes. They did divide them; that is, they considered ratios. But their understanding of division was limited to ratios between magnitudes of the same type. For example, they could deal with ratios between lengths and ratios between areas And they even knew what it meant to equate a ratio of lengths to a ratio of areas.

But the idea of an area divided by a length made no sense to them. All of their discoveries relating magnitudes of different types were expressed as relationships between ratios. Thus, Archimedes expressed his celebrated law of levers as a relationship between a ratio of lengths and a ratio of weights.[24] (See Chapter 7 for further details.)

That the classical Greeks could offer a rigorous definition of ratio, even to this limited extent, was a tremendous achievement. But such limitations are almost incomprehensible today. Today we routinely divide distance by time to calculate velocity. We multiply lengths to measure area and we multiply the number of years spent on a task by the number of people engaged in that task, expressing the result in man-years. As a matter of arithmetic we multiply and divide numbers, at will, keeping track of the units attached to each number.

But what relationships are we expressing, with such facility, by means of these calculations? What is the underlying reality, the actual relationships among magnitudes that we capture with our arithmetic calculations?

This modern facility, that we take for granted, was not easily won. As Harriman points out the modern paradigm was not available to Galileo.[25] At the dawn of modern science, Galileo was limited to the classical Greek paradigm, which left no way to express such fundamental concepts as miles per hour.

The ability to meaningfully multiply and divide magnitudes is a fundamental conceptual underpinning of modern science. But what are we doing when we multiply or divide magnitudes? What do these operations actually mean? I begin with the measurement of area.

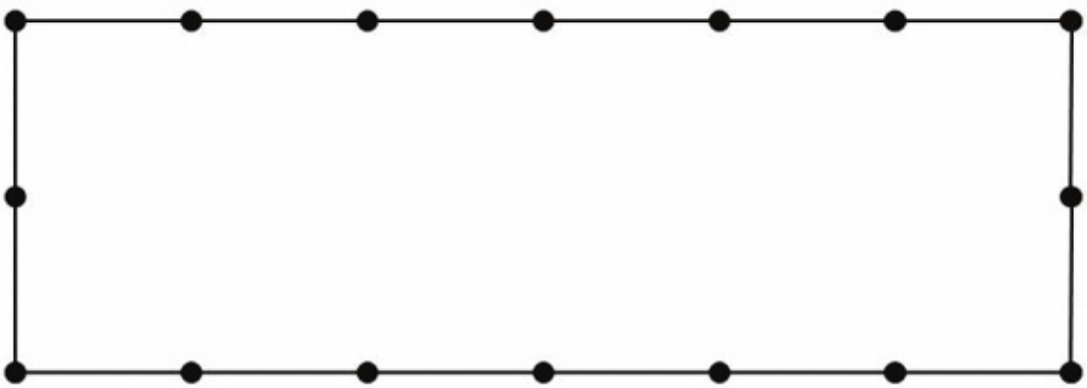
Area

Today, measurement of areas usually begins with rectangles. One says that a rectangle with a width of six units and a height of two units has an area of 12 *square* units. But a number of questions present themselves. First, where does this definition come from? Second, does it make sense? If one gets the same

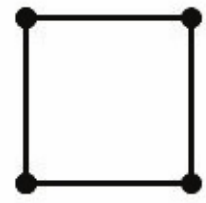
answer, for example 12 square units, for two different rectangles do they *really* enclose the same amount of area?

Finally, there is a third, more subtle question. What happens if one changes the units used to measure the two sides? If area is indeed a magnitude, the relationship between two different areas should be independent of the units one uses to measure the two areas. If one area, area X, is three times another area, area Y, then the *numerical* measurement of area X should be three times the numerical measurement of area Y, no matter what units one usea to measure the two areas, providing only that one measures the two areas in the same units whenever one compares them.

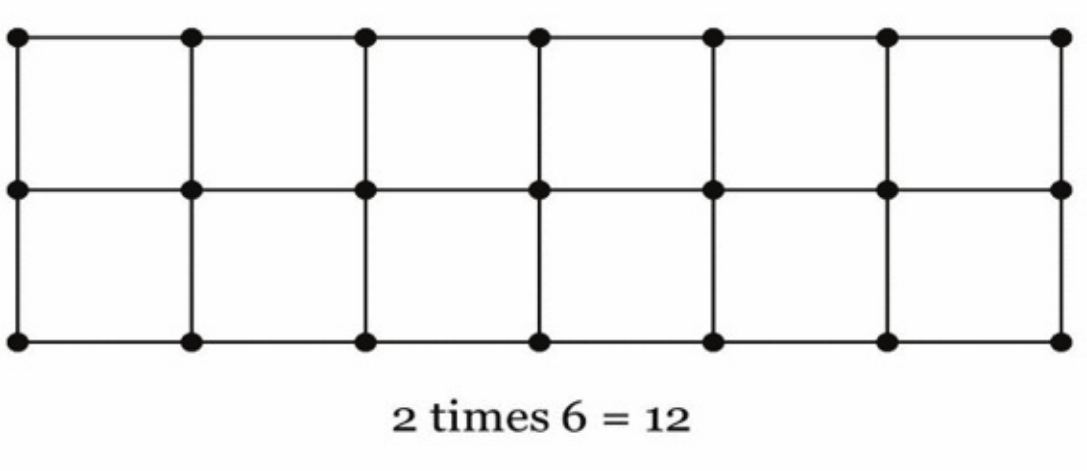
Consider the following rectangle:



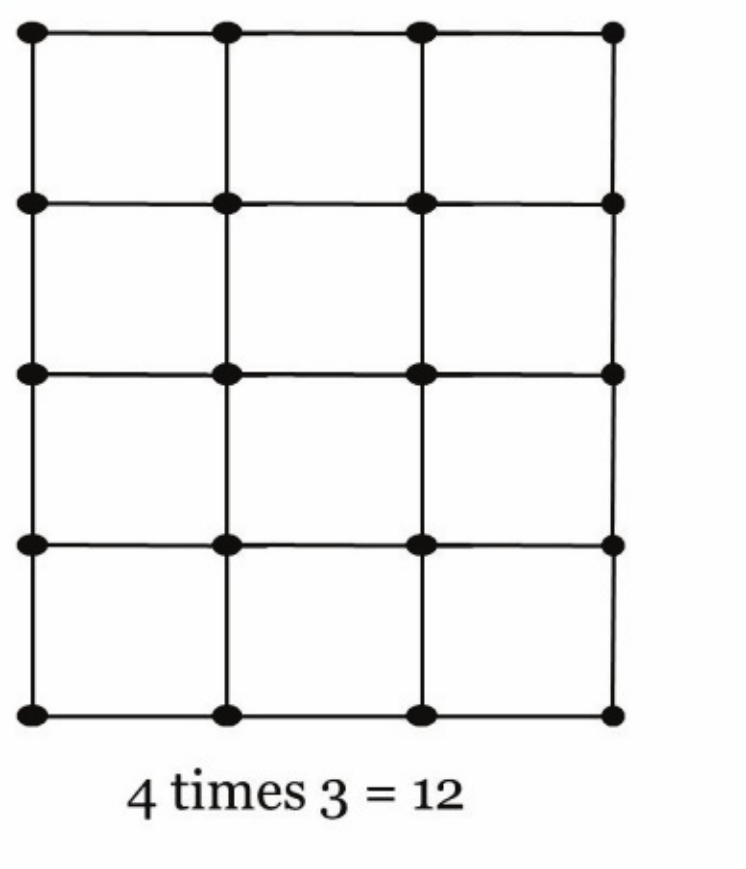
This rectangle is two units high and six units wide. Twelve square units! What is a square unit?
Well, what about:



A *square unit* is a square for which each side is of unit length. So now one just counts the squares:



A measurement in square units is just a count of the number of squares made out of units. So it now becomes almost obvious why a rectangle that is four units high and three units wide encloses the same area. *Because* 4 times 3 = 12 = 2 times 6, the new rectangle contains the same number of squares, literally, the same number of *square units*:

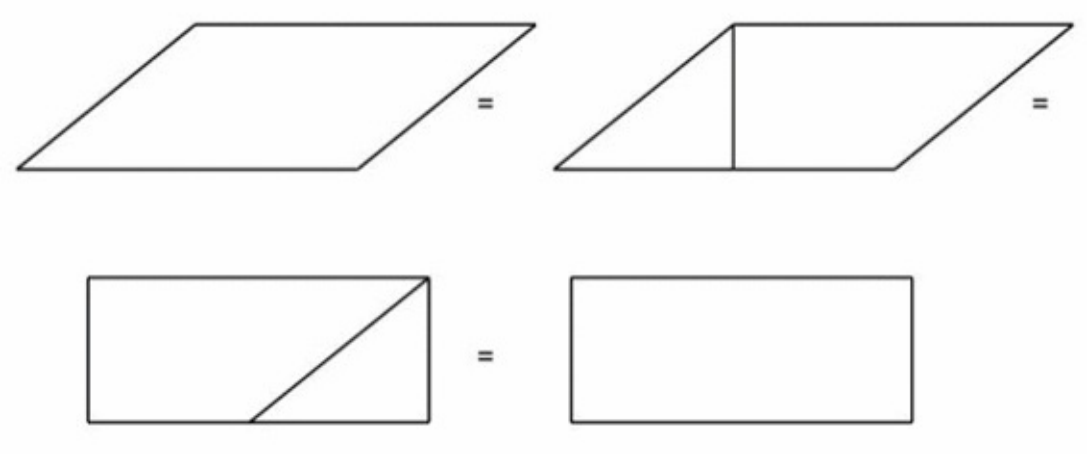


So the unit of area is a square and each side of the square unit is the magnitude

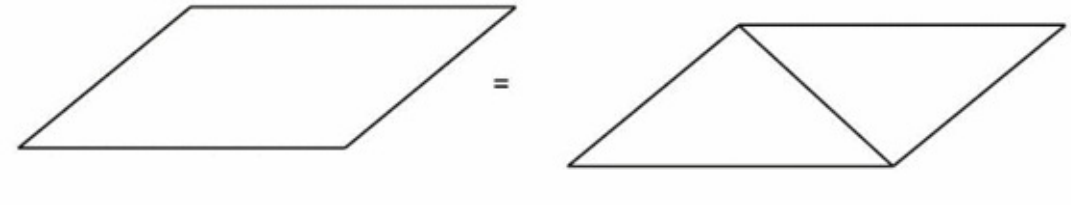
already in use as a unit for the sides.
The final question remains. Suppose that one changes one's choice of unit. Suppose, for example, that area A, as measured in square yards, is 2 square yards and that area B is 8 square yards. This means that area A contains 2 squares, each a yard long in each direction, and that area B contains 8 squares, each a yard long in each direction. Area B contains 4 times the area as area A.
Now suppose one measured in feet. There are 3 feet in a yard. A square yard contains an array of square-foot squares, an array that contains three units in both directions, a total of 9 square feet. Since each square yard contains 9 square feet, one converts square yards to square feet by multiplying by nine. 9 times 2 is 18 so area A is 18 square feet. 9 times 8 is 72 so area B is 72 square feet. To obtain the relationship between area B and area A, divide 72/18 = 4. So, by this calculation, area B still computes as 4 times the area of area A. The math had to come out this way because the factor that was used to convert A to square feet, namely 9 was also used to convert B to square feet. Multiplying numerator and denominator of a fraction by the same number does not change the result. Symbolically:

4 = (8 square yards)/(2 square yards)
= ((8 square yards) × (9 square feet per square yard))/((2 square yards) × (9 square feet per square yard)) = (72 square feet)/(18 square feet) = 4

This calculation is completely general. So the calculated ratio of areas is independent of the chosen unit of measurement.
From this, one could proceed to discuss squares with fractional numbers of units on its side. I will not pursue this here. But it is worth pausing to see how one can derive formulas for parallelograms, triangles and other figures simply by noticing various ways to subdivide them. For example the following picture illustrates that a parallelogram has the same area as a rectangle with the same base and height:



Similarly, the picture below illustrates that a triangle has half the area as a parallelogram with the same base and height:



Thus, one sees the familiar formula, that the area of a triangle equals one half of the product of its base and height.

To summarize, the product of two lengths provides a count of the number of the, possibly fractional, number of square units contained in a rectangle with those dimensions. The product of a length times a length is: a measure of area. When one multiplies two lengths, the result is *not* a length; but an area. When one multiplies two magnitudes, one does *not* get a magnitude of the same kind; one gets a square magnitude.

But Why Area?

What should one make of this multiplication of two line segments? One multiplies the width and the height of a rectangle and gets the area of the rectangle. But why rectangles? What if one were to multiply the two legs of a *right triangle*? Or what if one were to square the radius of a *circle*? Or multiply two sides of an *equilateral triangle*? Are such multiplications also measures of area? Do they measure anything about their respective geometric figures? And, if one interprets the answer as square feet … well, what does that interpretation have to do with the geometric figure in question?

Take the first case. Recall that a right triangle is a triangle in which two sides (the legs) of the triangle intersect at right angles. One knows that two right triangles can be put together to form a rectangle. So, therefore, a right triangle is half of a rectangle and has half of its area. So, from a modern perspective, one says that the *product of the legs* of a right triangle is *twice* the area of the triangle. In sum, it still makes sense to think of this product as an area: That product just has a different relationship to the area of the triangle than it does to the area of a rectangle. Instead of being the area of the triangle, it is twice the area of the triangle. And it makes sense to think of this product as the area that would result if one filled out the rectangle, containing the triangle, with the legs serving as base and height. In so doing, one creates a rectangle to which one can compare the area of the triangle, a rectangle with twice the area of the triangle.

The case of a circle is similar, but the relationship is less obvious. The area of a circle is not equal to its radius squared. But it is proportional to it. The area of a circle is proportional to a square with each side equal to the radius of a circle, a discovery known to Euclid. But Euclid did not know, and it was left to Archimedes to discover, that the proportionalityfactoristheconstant π,defined as the ratio of the circumference of a circle to its diameter. Squaring the radius is the first step to computing its area. That computation is completed by multiplying that square by the number π.

The final case of an equilateral triangle is also less obvious. But even in this case, like the case of a circle, the product of these sides is proportional to the area. And if Euclid were to establish that proportion, he would draw a number of *rectangles*, of various dimensions, and the relationship that Euclid would establish would be a relationship among *rectangles*. As an example of this sort of thing, Euclid followed this procedure in his arguments for the Pythagorean Theorem and for its later generalization.[26]

So the product, if interpreted as square feet, is proportional to the areas of each of these respective geometric figures and, in that sense, can be regarded as a measure of their respective areas. But the question remains: Why do we measure area in squares? And why do we interpret the multiplication of two lengths as a number of square units? Why not triangular units instead?

The first important thing to realize is that defining the *product* of lengths in the way I did involved a *choice*. I ask these questions not to imply that what we are doing is wrong, but simply to point out this element of choice. One *chooses*,

though for many good reasons, to measure area in squares; one does not measure area in, say, right triangles.

The practice of measuring area in relation to squares and rectangles goes back to antiquity. Euclid does not multiply lengths, as we do. But, as I discuss in Chapter 3, he does measure area, though in purely geometric terms, in relation to rectangles and squares. The Greek approach, however, was to find a square of the same area as the given figure, whereas the modern approach is to count squares of a given size. Thus, whereas we write a formula for the area of a circle, the Greeks spoke of "squaring" the circle, i.e., finding a square with the same area as a circle.

The modern approach certainly makes a natural choice, arguably the best choice, and it is so automatic today that it is hard to conceive an alternative. Yet there is an alternative, however inferior: in measuring area, we could have counted right triangles, say, with each leg equal in length to 1 foot, instead. If, perhaps, less intuitive, the result would constitute a measure, or a specification, of area in, one might say, "Triangle feet" instead of "square feet"".

As long as one retains the units involved, the product of two magnitudes expresses *something* determinate about the two magnitudes involved in the product. (I will elaborate this point later.) But to identify this product with a third magnitude, even a related magnitude, involves a choice. To multiply lengths to get an area, I identified that product with a *particular* area, namely the area of a particular rectangle, the rectangle determined by the lengths in the product.

But what, for example, is the product of two pencil lengths? There are two defensible answers. First, on the basis of my discussion, one might answer that it's the area of a rectangle with a length of the first pencil length and width of the other pencil length. That answer would presuppose that one has already interpreted the product of two lengths to be the area of a rectangle determined by those lengths, extending that interpretation to pencils.

However, I favor a different answer, namely that the question is inappropriate in the context of pencil lengths. The concept of area arises in, and pertains to, a specific geometric context involving the measurement of shapes. Comparing pencils, looking to multiply their lengths, is not part of that context.

The identification of a product of lengths with an area is a choice, but one that is

far from arbitrary. One has a limited array of choices and the availability of those choices needed to be discovered. One *discovers* that *this* magnitude, the area of a rectangle, relates in a certain way to the lengths of the sides of that rectangle: two *other*, though related, magnitudes. And, for this purpose, it's not even sufficient that the area be *determined* by the width and the length. It must be determined in a special way. For, by the nature of numerical multiplication, if one doubles either one of the two factors, for example, one doubles the product. If this product is to measure area, then the related area must double, as well. The area measured by the product of these two legs must be proportional to each dimension taken separately. It is *essential* that doubling the length or, *alternatively*, doubling the width, will double the area.

On the basis of this discovery , one *defines* multiplication of lengths, as a magnitude, as the *area* of a *particular* rectangle, namely the rectangle spanned by those lengths. And this definition, taking rectangles as a standard, also provides a way to measure the areas of other geometric figures, including circles and triangles. One relates the areas of each to the area of a square, one's chosen standard of area.

Once units of length are established, one, further, *establishes a relationship* between the unit that is used to measure lengths and the unit that is used to measure area. *On the basis of the geometric definition of multiplication for lengths*, one can relate the unit of length to the unit of area. A square with both sides equal to the unit of length has the *area* of a square with both sides equal to that unit of length. If one takes this "square unit" as the unit of area, then, as I've shown, the number of units of area will be equal to the product of the number of units of length and the number of units of width. In this way, one's choice of units is consistent with one's conception of multiplication, as applied to length. On the basis, and *only on the basis of this choice of unit* to measure area, one finds that 3 feet times 4 feet = 12 square feet. This choice of a unit of area guarantees that the multiplication of the *numbers* 3 and 4 will give the answer that corresponds to multiplication of the *magnitudes*, 3 feet and 4 feet.

Finally, keep in mind that multiplication of lengths is a way of *formulating* a relationship among magnitudes that existed *independently* of this formulation. The area of a rectangle has a determinate relationship to the lengths of its sides, independently of how one formulates that relationship. The *proportionality* of the area to either side, taken separately, is a *discovery*. Taken together, these facts *make it possible* to define an area as a *product* of two lengths. However, the

choice to do so is based on one further consideration: the fundamentality of area as a geometric magnitude and the use of rectangles to measure it.[27] That choice *recognizes* a fundamental relationship among magnitudes. Far from being arbitrary, it is almost necessitated by the facts it embraces. Nonetheless, it *is* a choice.

Multiplication of lengths to find an area is the simplest example of multiplication of magnitudes. It exemplifies, and is part of, a more general pattern: the subject of the next section.

Products of Other Magnitudes Newton's Second Law

The degree of acceleration of an object produced by an applied force on that object, depends on the mass of the object. Specifically, the *force* that is required to produce a particular acceleration is *proportional to the mass of the object*. Suppose for example, that the force is produced by a spring. If one combines two objects of the same mass, then the force produces half of the acceleration on the combined objects, compared to the acceleration that it would produce on just *one* of the two objects.

Secondly, the *acceleration* produced by a force on a particular object, an object of a particular mass, is *proportional to the force*. The combined action of *two* identical springs will produce *twice* the acceleration on a particular object as either spring would produce separately.

In sum, that required force is jointly proportional to both the *mass* on which it acts and to the *acceleration* that it produces. If mass is held constant, the required force is proportional to the acceleration it produces; if the acceleration is held constant, the required force is proportional to the mass on which it acts.

If one knows the mass of the object and the acceleration produced by a force, one can compute the amount of force required to produce the acceleration from the formula known as Newton's second law of motion. Newton's law unites three magnitudes that are connected in a particular physical context and formulates a *causal law* pertaining to that context.

The formula *presupposes* such a context; it would, for example, make no sense whatever to produce a formula relating the acceleration of a falling stone to the

force that someone exerts on the pedals of a bicycle. On the contrary, the causal relationship involves *one object* to which the force is *applied*. The mass in the formula and the acceleration in the formula both pertain to that object. The *force* is produced by an external agent *acting* on that object.

The second law of motion was an epochal discovery and is one of the most fundamental physical laws of nature.
One expresses Newton's law as force = mass times acceleration or, in symbols,

F = ma

The magnitude of the force is the product of the magnitude of the mass times the magnitude of the acceleration. "Force equals mass times acceleration" expresses a relationship among physical magnitudes.

But what is that relationship? The formula expresses a relationship, but, as it relates to any particular mass and acceleration, what does this formula actually mean? What is the amount of force required to accelerate a mass of 100 grams at a rate of 15 feet per second per second? How would one apply Newton's formula, or can one, based solely on the information so far presented? There is something, so far, missing from this discussion.

The analysis of force goes back to Archimedes. Archimedes knew nothing of the *dynamic* properties of force. But his investigations into mechanical levers and into the buoyancy of water were the first studies of *forces in equilibrium*.[28]

And these investigations made it possible to compare and measure forces long before anyone understood their effects on motion. One measures the force of gravity on an object by weighing it; one measures the force of a spring by comparing it to the gravitational force on a separate object. For this, Archimedes paved the way: One quantifies force, reducing its magnitude to perceptual terms, as the weight of a particular volume of a particular kind of object, a weight that one can experience directly by holding it in one's hand. One *measures* forces by *comparing* them to a chosen standard; one *compares* forces by bringing them into *equilibrium*.

But what is the *force* required to accelerate a mass of 100 grams at a rate of 15 feet per second per second? And what, as a second example, is the *acceleration* produced by a force of 60

pounds on a mass of 20 grams? How does the acceleration in feet per second per second produced by a force on a particular object relate to the measurement of that force in pounds? The force required to produce that acceleration is certainly determinate. And one can *specify* the force in relation to its dynamical effects. But *quantifying* that force in *pounds*, or in units that Archimedes would have recognized, requires an act of discovery. To multiply a mass, specified in grams, by an acceleration, specified in feet per second per second, and then expressing the results of the multiplication in gram-feet per second per second, is not, *per se*, to identify how many *pounds* of force was required to produce that acceleration.

Newton's Law expresses a *causal* relationship relating physical quantities. Once meaningful units have been identified, mathematics can *express* the causal relationship in relation to these units. But the mathematics cannot shortcut the process. Only physical experiment can, establish that relationship, can, in particular, establish the physical units required to relate the arithmetic calculation to the reality that it is intended to express. In sum, if all I knew was this formula, if I had no independent knowledge about how any *particular* mass and any *particular* acceleration related to the particular physical force required to *produce* the acceleration, I would be unable to apply it to any specific case. The formula pertains to an existing causal relationship and it provides information about the nature of the relationship, but doesn't, all by itself, completely *capture* that relationship. From the formula, one knows that the force is *proportional* to the mass and the acceleration it produces; but that knowledge, by itself, *does not suffice to measure that force, as an equivalent weight, in pounds or ounces*.

The formula needs to be *calibrated* somehow. Each magnitude can be measured separately, but the only way to calibrate the *magnitude* of the force representing the product of mass and acceleration is to perform an experiment. Only when one knows, for example, that the *force* required to produce some particular known acceleration on an object of some particular known mass *would balance a weight* of, say, 200 pounds on a balance scale, can one calibrate the formula. From a mathematical perspective, one experiment is enough to calibrate the formula. And the way it's usually done is to first choose a unit for acceleration, such as meters per second per second, and a unit for mass, such as kilograms. And then determine the *particular* force required to produce a unit of acceleration (one meter per second per second) on one unit of mass (one kilogram). Having determined that force experimentally, *as equal to the weight of a standard object* (an object specified, for example, as a particular volume of water at a particular temperature, at sea level), one can define that

experimentally measured force as a new unit, a unit relatable to the older standard.[29]

As in the case of area, this kind of choice guarantees that the formula between lengths will work for numbers, as long as the numbers correspond to the measurement of lengths in the appropriate units. In the case of Newton's law, one defines a unit of force, appropriately named the "newton", defined as "the amount of force required to accelerate a mass of one kilogram at a rate of one meter per second per second."[30]

But notice a couple of things about this choice of units. First, notice that the newton is not specifically defined in relation to weight, whether in pounds of kilograms. Rather, it is expressed in terms of the other units involved in the formula. So, if one measures mass in kilograms and acceleration in meters per second per second, then for m = 1 and a = 1, this choice of unit for force guarantees that F = 1. Any other choice of units for F would result in a different numerical answer for F (expressing, however, the same relationship, among physical magnitudes). So F = ma in the metric system, only with this choice of units to measure force; any other choice would require a conversion factor to convert the numerical answer to newtons. The formula, in its usual expression is valid numerically, in other words, *only* when the various physical units are chosen, in this way, to make the numbers correspond properly, to make the number of mass units times the number of acceleration units equal the number of force units.

But this leads to the second point. In defining newton in this way the question of relating the force represented to the weight of some standard object under some standard set of conditions is left unanswered. That question has been neatly finessed and the final connection to the world, the required calibration comparing a newton to the weight of a standard object, still remains to be made. Now, of course, the required experiment was performed long ago, long before the newton was defined, and was part of the context within which the newton was defined. So my point is not that there is anything problematic about any of this. Rather it is simply to emphasize that a quantitative relationship among physical magnitudes, expressed mathematically, is not, strictly speaking, a mathematical relationship. It is, rather, irreducibly, a mathematical *expression* of a physical relationship.

In other words, the meaning of F = ma is not determined by the mathematics; it is determined by the physics and requires physical experiments, not only to establish the type of relationship, but to *calibrate* that relationship. The relationship between the variables is not driven by mathematics; it is not a

specifically mathematical relationship, independent of physical context. Rather, it is a quantitative relationship pertaining to, relating quantities involved in, a particular kind of physical context. It is a quantitative relationship that quantitatively expresses and identifies a *causal* relationship.

I have now discussed two examples of multiplication of magnitudes. One arises in a geometric context and the other arises as an expression of a causal, physical law. Both cases provide a natural choice of unit for the product, one that corresponds to the standard units used for the two magnitudes that have been multiplied. In both cases, reality sets the calibration and the meaning of the multiplication operation. A particular area, the area of a particular rectangle, is the area that corresponds to the product of the lengths of its sides. A particular force, whatever force be, as a matter of fact, required to produce a particular acceleration on a particular mass, is the force expressed by the product of mass and acceleration. The mathematics does not create these relationships. On the contrary, it is used to *express* them.

Products and Units

Let's look at this more generally.

It is only possible to establish or meaningfully define a numerical relationship of a product of two magnitudes to a third magnitude, under certain conditions. First, there must be a third magnitude that is determined, in some context, by the two magnitudes. Secondly, as in the cases of area and Newton's law, the relationship to this magnitude must satisfy a very special condition. The third magnitude must be proportional, in the nature of things, to each of the two factors, taken separately. For example, if one doubles either magnitude in the product, one doubles the result. If one multiplies either magnitude involved in a product, by a number, one multiplies the resulting product by the same number. Multiplication must work the same way with magnitudes as it does with numbers: As a representative numerical example, if 4 times 5 is 20, then 8 times five is twice 20, 8 being twice 4.

A magnitude can, certainly, be related to two other magnitudes without satisfying the proportionality condition. But then that relationship cannot be expressed as a multiplication of two magnitudes.

Suppose that, in some physical setting, a magnitude Z is determined by, causally related to, two other magnitudes X and Y. In mathematical terminology, Z is a function of X and Y. Any pair of magnitudes X and Y determines a specific magnitude Z. One writes $Z = f(X, Y)$ to designate this relationship. Again, in this

expression, X, Y and Z are *magnitudes* (*not* numbers) and *f* is the *name* that I am giving the *relationship*, whatever that relationship might happen to be. So, in this expression, X and Y are the magnitudes that determine, physically, the magnitude Z and f(X,Y) means "the physical magnitude Z that corresponds to magnitudes X and Y".
For the relationship *f* to be definable as a *product* of magnitudes, the value of Z must be proportional to both X and Y, taken separately. For example, if X is doubled and Y remains unchanged, then Z will double. In my notation,

$f(2\times X,Y) = 2\times Z = 2\times f(X, Y)$

In this formula, interpret $2\times X$ as I did in my discussion of the prearithmetic of magnitudes. 2 times X is simply X + X. In words, if X is replaced by 2 times X then Z is replaced by 2 times Z.

By the same token, if one multiplies Y times 3 and leave X unchanged, the value of Z also triples. In my notation, this means $f(X, 3\times Y) = 3\times Z = 3\times f(X, Y)$.

In general, using A and B to represent any numbers whatever,
$f(A\times X, B\times Y) = B\times f(A\times X,Y) = A\times B\times f(X,Y) = A\times B\times Z$

This formula expresses, in symbolic terms, that Z is proportional to either X or Y taken separately.
Now select physically defined units for magnitudes X and Y, and designate them as U_X and U_Y, respectively. (I use the subscripts in my symbolism to keep track of the respective magnitudes that each measures.) Remember that a unit is just a *particular* magnitude of a particular kind that has been chosen as the *standard*. Keep in mind that there is no reason to simply assume, in advance, that the unit used to measure Z has anything to do with the units used to measure X and Y. By the nature of the case, Z is a different kind of magnitude than X and Y and the Z-type of magnitude may have arisen or been discovered in another context (as was the case of force in the previous example).
One does not choose the particular physical *magnitude* Z that corresponds to specific magnitudes X and Y. Reality makes this choice for us, but one does have a choice of *units* that one uses to *measure* Z. So suppose one *chooses*, as a unit for the magnitude Z, the value $U_Z = f(U_X, U_Y)$. This choice takes U_Z, as a unit for measuring Z, to be the *particular* magnitude of that type that, as it happens, relates to the *particular* physical magnitudes U_X and U_Y.
What happens if one expresses X, Y and Z in terms of these units? Each is some

particular multiple of its respective particular unit. So there are numbers A, B, and C such that $X = A \times U_X$, $Y = B \times U_Y$, and $Z = C \times U_Z$.
Recall that X, Y , and Z are related by the function *f*, that Z = f(X, Y). On substituting the expressions for each variable in terms of units, one finds:

$$C \times U_Z = f(A \times U_X , B \times U_Y) = A \times B \times f(U_X , U_Y) = A \times B \times U_Z$$

In other words, C = A×B. When expressed in these units, the relationship among the magnitudes is *expressed* by the arithmetical multiplication of *numbers*! But keep in mind that, as an expression of the physical relationship, the meaning of these numbers, A, B, and C is totally dependent upon the choices of units U_X , U_Y and U_Z, choices that are a necessary part of the context of the numerical formula.

The physical *relationship*, of course, does *not* depend upon the choice of units. And one does not always choose units to simplify the expression of some particular physical relationship. But sometimes one does when that relationship is a fundamental one, such as Newton's second law of motion.

In any particular context, one can make this choice of units just once for a pair of magnitudes of particular kinds. For example, the area of a triangle is directly proportional to the base and the height. But one would *never* say that the area of a triangle is *equal* to the product of the length of its base and its height – because the product of two lengths has already been *taken* to be the area of a *rectangle*. We do not measure area in triangle units. One says, rather, that the area of a triangle is *half* the product of its length and its base. In measuring triangles, one takes for granted that 1 foot multiplied by 1 foot equals 1 square foot, i.e., the area of a 1 foot by 1 foot square. (Or, if one changes the standard of length, one makes a corresponding change in the standard for area. For example, if length is expressed in meters instead of feet, then area is expressed in square meters.) Accordingly, in numerical calculation, this relationship of the units, that multiplication of lengths measures the area of a *rectangle*, is also taken for granted.

Returning to the magnitudes X, Y, and Z, what if one chooses a *different* set of units for measuring Z, such as the standard that was already in place before the discovery of the relationship *f*? Then one needs to find a conversion factor to calibrate the formula. How?

Take U_Z, *in this argument*, to be the *preexisting* standard for Z. Express X, Y, and Z in terms of their respective units. Once again, A, B, and C are *numbers* that relate each magnitude to its unit:

$X = A \times U_X$, $Y = B \times U_Y$, and $Z = C \times U_Z$

Starting from the given relationship of magnitudes ($Z = f(X, Y)$) and substituting the above expressions for each magnitude, we obtain:

$C \times U_Z = f(A \times U_X , B \times U_Y) = A \times B \times f(U_X , U_Y)$ so $C \times U_Z = A \times B \times f(U_X , U_Y)$

But $f(U_X , U_Y)$ is a magnitude. It is, in fact, the same kind of magnitude that Z is and that U_Z is. $f(U_X , U_Y)$ is the particular magnitude of type Z that is determined in the prescribed way by the magnitudes U_X and U_Y.

So, for some *number* L, determined experimentally, one can express $f(U_X , U_Y)$ in terms of U_Z as
$f(U_X , U_Y) = L \times U_Z$

The number, *L*, captures the relationship between the magnitudes $f(U_X , U_Y)$ and U_Z , a relationship that the Greeks treated as a ratio. In general, if R and S represent magnitudes of the same type, there is a *number* D such that $R = D \times S$. As discussed earlier in the section on the Axiom of Archimedes, I define D to be the *ratio of R to S* and write *R/S = D* as simply an *alternate way* of expressing the relationship between R and S. (As we discussed earlier, D, in a physical application, is defined up to materiality.) R/S is a notation to express that relationship, that ratio, and its value is D. A *ratio*, then, is a number that designates the multiplicative relationship between two magnitudes (or multitudes) of the same kind.

In this terminology, one can rewrite this expression as
$f(U_X , U_Y)/U_Z = L$
L is the ratio between the two magnitudes of type Z. Substituting for $f(U_X , U_Y)$, one finds,
$C \times U_Z = A \times B \times L \times U_Z$
from which it follows that
$C = A \times B \times L = L \times A \times B$

The arithmetic formula $C = A \times B$, that resulted from my earlier choice of unit, has been replaced by $C = L \times A \times B$. In this expression, the coefficient L is the *conversion factor* to convert numerical measurements that are expressed in units of $f(U_X , U_Y)$ to numerical measurements that are expressed in units of U_Z. The arithmetic expression has changed; it is affected by the choice of units. But the relationship among *magnitudes* that it *expresses* remains the same and, indeed, the derivation of L presupposed and relied on this preexisting relationship among magnitudes.

Products: General Case

The need to multiply magnitudes is not limited to cases in which their product is a magnitude that has been previously or independently identified. Take, for example, the measurement of momentum. For purposes of this discussion, consider the onedimensional case in which all velocities run along the same direction.

The *momentum* of an object is measured as the product of its mass and its velocity. For example, suppose an object has a mass of 50 kilograms and moves at a speed of 100 feet per minute. Then its momentum is 5000 kilogram-feet per minute.

What about this product?
First, it is a physical fact that the object has a mass of 50 kilograms and a speed of 100 feet per minute. Secondly, it is a fact that if the mass is expressed in kilograms and the speed is expressed in feet per second, that the *arithmetical* product of these measurements is 5000. One acknowledges the context of this multiplication by keeping track of the units involved in the measurement of the magnitudes participating in the multiplication.
So the result of the calculation reflects something *specific* about the object to which it pertains. One can think of it as capturing the amount of ongoing motion of the object, capturing it from a perspective that regards the amount of motion involved, say, in *two* such moving objects as representing *twice* the amount of ongoing motion as the motion of either object taken separately.
But, unlike the case of Newton's Second Law, there is no independent measure of momentum to which this measurement of kilogram feet per minute need relate. The calibration requirement does not arise. It cannot arise unless and until one would discover some other manifestation of an object's momentum, a

manifestation giving rise to a different way to compare momenta and giving rise to an independently identified standard of measurement.

The need to relate two, separately measurable, physical manifestations of force is the element that was present in the case of Newton's Law and that is missing in the present case. The need, in the case of Newton's Second Law, to relate newtons to pounds was a need to integrate all of the relevant known facts pertaining to the measurement of force.

Yet, as I have already argued, momentum is a specific property of a moving object. It is a property; indeed it is a magnitude, a magnitude that can distinguish one moving object from another. But why does one care?

Notice, first of all, what is being lost when one multiplies mass times speed. An object with a mass of six kilograms and moving at a speed of 50 feet per minute, for example, has the *same* momentum as a second object with a mass of two kilograms moving at a speed of 150 feet per minute, namely, 300 kilogram-feet per minute. The two objects have different masses, different speeds, but the same momentum.

When one focuses on momentum, neither mass nor the speed, taken separately and in isolation, is important. Only the product matters. The *product* of the two factors is everything, the only thing that matters; the specific *composition of the two factors* is treated as an omitted measurement.

One can, as I have indicated, simply regard momentum as *a* measure of the amount of ongoing motion. But, as it happens, momentum is of great importance. It is important because the devastation caused by a collision with a moving object primarily relates to momentum. It is much worse to be hit by a truck going 25 miles per hour than by a Volkswagen Beetle traveling at the same speed and far worse to be struck by a train. And these physical consequences reflect a fundamental principle in physics, namely the conservation of momentum. As Resnick and Halliday formulate the conservation of momentum principle, "When the resultant external force acting on a system is zero, the total vector momentum of the system remains constant."[31]

One cares about momentum, in large part, because of this conservation law and the consequences of the conservation law. But to *formulate* that law, to *discover* it in the first place, one needs to be able to *meaningfully multiply two physical magnitudes and express the results in terms of the chosen physical units of the two factors*, prior to such discoveries and formulations.

To further explore this point, as a final example, consider a balance beam with a weight on either side of the fulcrum. If the weight on one side is six pounds and its distance from the fulcrum is ten inches, the product, 60 poundinches, is called the *moment* of that particular weight with respect to the fulcrum of the lever.

That six pound weight will be balanced by a three pound weight situated 20 inches from the fulcrum on the other side of the lever. And it will balance precisely *because* the moment of this second weight, calculated as three pounds times 20 inches, is *also* 60 poundinches.

In this example the meaning and importance of this product of magnitudes is entirely bound up in a particular physical context, namely that of the lever arm. Yet the calculation of each moment reflects a specific precise characteristic of the physical situation of each weight, has a specific meaning and import within that context, and is important because of that context.

In general, a product of magnitudes is a *condensation* of certain aspects of a physical situation. The product reflects the measurements of each factor, but it also omits certain characteristics of the physical context, retaining only the arithmetic *product* and the *combination of physical units* that relate this product to the physical context of the measurement. In relation to these units, the product measures a physical characteristic of the physical situation to which it relates. And this product is properly regarded, in its own right, as a *magnitude*. Indeed, one sees that the product is a magnitude simply by holding one of the factors constant and observing that any arithmetical operation involving the second factor applies immediately to the product, as well.

In general, whenever two magnitudes are connected within a category of physical situations, their product can be taken as a third, derived magnitude pertaining to and measuring something about that type of physical situation. Thus, for example, the momenta, of objects moving in the same direction, can be compared, added, and related as to multiplicity. The law of conservation of momentum, in particular, *presupposes* such quantitative relationships, *presupposes* that momentum can be regarded as a physically meaningful magnitude, subject to further investigation. The ontological status of momentum as a magnitude is not affected by the indirect means required for its specification and measurement.

Division of Magnitudes by Magnitudes

The ability to calculate physical quantities such as speed, acceleration and density, to generally relate magnitudes of different types, is an essential underpinning of modern science. But the ability to make these calculations is a modern development; as I have indicated, and except for taking ratios among magnitudes of the same type, this ability was unknown to the ancient Greeks.

Yet the underlying *rationale* for these ratios is seldom, if ever, acknowledged or discussed. To provide such a rationale, to even recognize that one is needed, cannot be found in the standard curriculum, not in mathematics, nor in physics, nor in the philosophy of science. We take it entirely for granted. But our warrant for doing so is not obvious and merits investigation.

Density

I begin with density because it provides an ideal example of the essential issues. Density pertains to solid objects, liquids, and confined gases. Density is a measurement of the amount of matter, the amount of mass, contained in a particular volume. If the object, the liquid, or the gas is homogeneous, then, other things being equal, the density is constant – independent of the particular volume under consideration. Thus, 20 grams contained in a volume of 2 cubic inches represents the same density as 10 grams contained within one cubic inch. Density is regarded as constant when the amount of mass contained within a volume of material is proportional to the volume under consideration.

For example, the density of gold at a particular temperature is characteristic of that element and is totally independent of the size and shape of a particular sample. Or, to take another example, the density of a confined gas is normally constant throughout the container.

How does one measure density? Density is a magnitude so one establishes a relationship to a standard, to a particular concrete. But how does one characterize such a concrete?

One could, perhaps, select a particular dense object. For example, one might select gold, at 72^0, or ice, just below its melting point. But that is not what one normally does. And the reason is that density does not admit of direct comparison; one requires indirect means to compare the density of two objects. Density is a measure of the amount of mass in a particular volume so, to determine density, one must determine the mass and the volume.

Taking weight as an indicator of mass, if one object has twice the weight of another object, having the same size and shape, it is twice as dense: It has twice the mass in a particular volume. For a particular volume, the density is directly proportional to the weight.

On the other hand, if two objects have the same weight, but one has half the volume as the other, then the smaller object has twice the density. If the larger volume be divided in two, each of those two halves will match the volume, but contain half the mass, as the smaller object. For a particular weight, density is

inversely proportional to the volume containing that weight.

In general, two objects with the same ratio of weight to volume have the same density. For example, suppose object A weighs 16 ounces (oz) and has a volume of 2 cubic inches (in^3), while object B weighs 24 ounces and has a volume of 3 cubic inches. Then one cubic inch of object A will weigh 8 ounces, since one cubic inch is half of 2 cubic inches: *I have divided the volume by 2 so I need to divide the weight by 2*, since, by assumption, the weight is evenly distributed throughout the volume. One finds 16 oz/2 = 8 oz. Similarly, one cubic inch of object B will contain one third of the weight of object B, namely 24 oz/3 = 8 oz. So objects A and B have the same density. Two objects with the same relationship between weight and volume, two objects that, when expressed in the same units have the same numerical quotient, have the same density.

Moreover, this numerical quotient provides a measurement of the density, considered as a magnitude: If the mass in a particular volume doubles, the magnitude of the numerator doubles *and*, therefore, the numerical quotient doubles. The quotient precisely captures the relationship of the density of an object to its mass and volume.

But what is the standard of measurement?

A constant density is independent of the *size* of the volume or of the *particular* volume of that size that one considers. So choose a standard volume, such as one cubic inch. If one knows the mass of a cubic inch of material, one knows the density of the material. If one measures mass in ounces, then the density of the material is measured in ounces per cubic inch. If there are, for example, five ounces in a cubic inch, than *any* cubic inch of the substance contains a mass of five ounces. The measurement of density answers the question: How many ounces of this material are there in a cubic inch?

If an object weighs 16 ounces and has a volume of 2 cubic inches then its density is 8 ounces per cubic inch *because* one cubic inch of this material will contain 8 ounces of it. The weight of any particular cubic inch of the material is direct *manifestation* of its density.

So, to find the density of a material, divide its mass by its volume and keep track of the units one uses for both the mass and the volume. In this case, one represents this quotient as

16 oz/2 in^3 = 16/2 oz/in^3 = 8 oz/in^3 or 8 ounces per cubic inch.

The identification of the units at the end designates the fact that, to measure density, one measures the mass, in ounces (the numerator) *within* a particular volume in cubic inches (the denominator). “Per” in this context means “for each”. In this example, it means that each cubic inch contains 8 ounces.

But notice something important. In performing this calculation, I have not *literally* divided 16 ounces by 2 cubic inches. I have *not actually* divided two magnitudes of different kinds. And this is indicated by the logic of my earlier example. What I actually did was divide 16 ounces by a number, a number specifically representing the relationship of the volume of the object to a unit of volume. I divided two cubic inches by one cubic inch to find the fraction of the whole that I needed to consider. I then divided the weight of the sample volume by the *ratio* of that volume to a unit volume to determine the weight of a unit volume.

In other words, I reasoned that 16 ounces per 2 cubic inches equates to 8 ounces per 1 cubic inch. The answer, 8, is half of 16 *because* 1 is half of 2. The answer, taken with the context supplied by the reminder of the units involved, and the way that they are involved (oz/in^3) is a complete specification of the density, one that, in fact and despite superficial appearance, does not require dividing one magnitude by a second magnitude of a different type.

And this is the general principle. Division of two related magnitudes can generally be resolved by relating the divisor, as a ratio, to its unit, dividing the numerator by the resulting ratio, and retaining the units.

For every division, there is a corresponding multiplication that reverses the division. What about this corresponding multiplication in the case of density? Keep in mind that density is a magnitude that relates, for a particular object, the mass contained within a volume to a measure of that volume. Density is calculated by dividing the mass, within the particular volume containing the mass, by the ratio of that volume to a standard volume.

The corresponding multiplication is straightforward: Density times volume equals mass. How much mass? The mass contained within the volume. If the density of a homogeneous substance is 5 grams per cubic inch, one finds the mass of a 10 cubic inch volume by multiplying 10 cubic inches by 5 grams per cubic inch. One takes this product as answering the question, “How many grams of material are contained within 10 cubic inches of the material?” And the answer to this question is found in exactly the way that one would expect: Multiply 10 times 5 and express the result in grams: 10 cubic inches time 5 grams per cubic inch equals 50 grams.

Speed

The corresponding analysis, in its entirety, applies to speed. Speed is a measure of the distance traversed within a specific time. If one travels, at a constant

speed, a distance of 300 miles in 5 hours, one must have traveled 60 miles during each of those 5 hours. One's speed, then, is 60 miles per hour. One divides the total distance traveled by the ratio of 5 hours to one hour. But this ratio is 5 and 300 miles divided by 5 is 60 miles. Since these 60 miles were traversed during the interval of one hour, the speed is 60 miles per hour.

Conversely, if one knows the speed and duration of the travel, one multiplies these two magnitudes to find the total distance traveled. Thus, 5 hours times 60 miles per hour is 300 miles.

So, in a sense, the Greeks were on the right track. One does *not*, appearances to the contrary notwithstanding, need to define a ratio of a magnitude to another magnitude of different type. It suffices to find the ratio of the denominator to its unit and, then, to divide the numerator by the resulting number.

Finally, I have made an assumption in each of the last two examples that needs to be relaxed. In the case of density, I assumed constant density. And, in the case of speed, I assumed constant speed. But the respective quotients are valid, even without such an assumption, providing that one properly reinterprets the meaning of these quotients.

In the case of density, suppose a material weighing 200 pounds and occupying a volume of 10 cubic feet. Without knowing the actual distribution of material within this volume, one can, nevertheless, say that the density of the material, *on the average*, is 20 pounds per cubic foot, meaning that *if* the mass were uniformly distributed, that would be its uniform density.

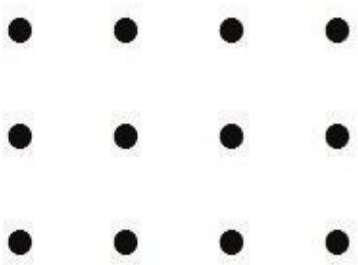
Similarly, if one travels 300 miles in 5 hours, then the *average* speed was 60 miles per hour, regardless of the number of times one might have slowed down, speeded up, or stopped altogether. Again, that average speed of 60 miles per hour is the speed that one would have gone *if* one had traversed that 300 miles at a constant speed during those 5 hours.

This notion of an average is also one key to understanding the concepts of instantaneous velocity at a particular time or density at a point. Instantaneous velocity, for example, is simply one's average velocity during a sufficiently small temporal duration at the time in question. Likewise, density at a point is the average density of a sufficiently small volume containing the point in question.

How small? That depends, as always, on one's precision requirement in any concrete instance.
When physicists perform calculations involving physical magnitudes, they focus on the physics and take entirely for granted their ability to multiply quantities or divide them whenever necessary. They simply make their measurements, perform their calculations, and keep track of the units and conversion factors involved.
The examples I have just given are representative of the standard approach: Express all magnitudes in terms of physical units; perform whatever arithmetical calculations are pertinent to one's purpose; and keep track of the units that were applied to the various magnitudes, distinguishing the units relating to the magnitudes in the numerator from those units relating to the denominator.
But one usually performs these steps without fully understanding just why it makes sense, without even an inkling that there is anything to think about, that the possibility of such calculations ever required discovery or that the rationale could ever have been in doubt.

Numbers and Units

So much for multiplying magnitudes! What about apples or oranges or dots? How should one analyze the following 3 by 4 array?



What multiplication does this picture represent? Does this picture illustrate multiplication of four dots by three dots? Or does it, rather, depict three *repetitions* of four dots?

Well, the total is 12 dots, not 12 *square* dots, so it must represent three repetitions of four dots. As I already had occasion to recall, multiplication is first taught as a short cut for repeated addition. From this perspective, one multiplies four dots by *three* and gets 12 dots. Here, three is not a number of dots, but a

number of repetitions – exactly as when one multiplies a line segment by a number to get another line segment. The two situations are essentially identical.

So, in this context, when one says that three times four equals four times three, one is saying that three repetitions of four *units* equals four repetitions of three *units*. When one looks at the rectangular array in two different ways to understand why these two products are equal, one is truly looking at it two *different* ways. In the one case one multiplies three dots; in the other four. The rectangular array represents the result of either process.

Now when one arranges numbers on a number line, it's as though these numbers had equal status, as though they represented different amounts of the same unit, as though they represented magnitudes of the same kind. One adds 2 and 3, numbers on the line, and gets 5, another number on the line. One multiplies 2 times 3, numbers on the line, and gets 6, another number on the line. One takes the square root of 25, a number on the line, and gets 5, a number on the line. Yet we have just seen that, in the case of multiplication, at least one of these factors would have to be interpreted as a number of repetitions to get an answer of the same kind of magnitude as the other factor.

So what's going on here? If one multiplies two instances of a unit, the result must be expressed in square units.
Well, perhaps the unit of the real line is, implicitly, "number of repetitions." One could certainly argue that 3 repetitions times 2 repetitions equal 6 repetitions. But this interpretation, or any other, must address one key fact: it shouldn't *matter* what the units are. Where, after all, did this abstraction of 'number' come from? And to what does it apply? Clearly, the laws of arithmetic should apply no matter what the units might be; no matter what kind of magnitude or multiplicity they are applied to. So, it would seem, we have hit on a problem!
Well, consider the following relationships:

● 3 times 4 feet equals 12 feet
● 3 feet times 4 feet equals 12 square feet ● times four dots equals 12 dots
● 4 times three dots equals 12 dots
● 4 miles per hour times 3 hours equals 12 miles

One can sum them all up as
● 3 of something times 4 of something equals 12 of something

In this formulation, the word, *something*, appears three times, but each time it appears, it might mean something else. But the numbers are always the same; the arithmetic does not depend upon the units to which it is applied.

No matter what "something" refers to, the same *arithmetical* fact underlies every one of these statements. Namely, 3 times 4 equals 12

I've now, finally, omitted any mention of units from the equation. To paraphrase Ayn Rand, the omitted units must be units of some kind, but they may be units of any kind.[32] When one multiplies a number of units of one kind by a number of units of a second kind, the result is a number of units of yet a third kind.

More remains to be said about the need to be *consistent* in ones choice of units between the various terms of the calculation. But, assuming such consistency for now, the *number* of units of this third kind depends only on the numbers. And this finally gets us to something very familiar: When one multiplies a number times a number, the result is a number.

Treating units as omitted measurements in just this way is the step one needs to take to get us to the number line. It is not that the units of the product are the same as the units of the factors; it is that when ones focus is on the *arithmetic* the particular units being omitted in that particular calculation don't matter. One ignores them because they are not relevant to the problem in arithmetic.

It is important, though to understand another, more fundamental, way to look at number. In essence, a number designates a *relationship*, a *ratio*, between a quantity and the unit used to measure the quantity. A number, as such, has no units; rather it stands for a *relationship* that is independent of unit. One can certainly look at number as involving omitted measurements, as one does with any concept. But, what is being omitted, (among other things) is the particular *kind* of magnitude or multitude being related. Fundamentally, a number simply stands for the *relationship*, independent of the specific choice of magnitude that it may be used to relate.[33]

And this brings us to the confusion that can be engendered by the number line. The number line carries the suggestion that a number is a kind of magnitude. But it isn't; it's a kind of *relationship*. The laws of arithmetic are laws that relate *numbers*. And numbers, in turn relate pairs of *magnitudes* or pairs of *multitudes*. Numbers relate a magnitude to a chosen standard and whole numbers relate a

multitude to a multitude of one.

In my treatment I have emphasized the relationships between magnitudes: the sum of two magnitudes, the product of a magnitude and a number, the product of two magnitudes, and the quotient of two magnitudes. These relationships are the basis of arithmetic. Numbers add the way they do because multitudes and magnitudes add the way that *they* do. The laws of arithmetic derive from the corresponding relationships among multitudes and magnitudes. But arithmetic is a higher order of abstraction and it is a mistake to confuse a number with either a magnitude or a multitude.

Descartes: Geometry and Arithmetic

This is not the answer that Descartes provides in his *Des matiers de la Geometrie*, though he faced similar issues.[34] Now, as we shall see, Descartes was asking a different question, so one should not expect the same answer. Yet he is sensitive to issues of this sort, first, because of his familiarity with Greek geometry, his point of departure. And secondly, he needed to address the issue of units because his purpose was to *reduce questions in geometry to questions in algebra*. Descartes's discussion is historically important and it will illuminate, and provide a foil for, my own observations.

Descartes' enterprise includes two basic steps. The first is the introduction of coordinates, the ancestor of the real number line and also the ancestor of the modern form in which we utilize Cartesian coordinates. The second step is the use of equations, such as $y = x^2$ to specify geometric shapes algebraically. This equation of a parabola is used to designate a graph consisting entirely of points, represented by pairs (x,y), that satisfy the equation. For example, x =3, together with y =9 satisfies the equation because 3 squared is 9.

Descartes shows his awareness of my issue at the very beginning of his *Geometrie* when he states:

"It should also be noted that all parts of a single line should always be expressed by the same number of dimensions, provided that unity is not determined by the conditions of the problem. Thus, a^3 contains as many dimensions as ab^2…"[35]

Descartes is saying, for example, that, without specifying a unit, the expression y

$= x^2 + x$ is meaningless because x^2 and x have different units. However, once a unit has been specified, it is possible to interpret the expression. Now, from my vantage point, I respond, “Yes, because then x and y can be treated as numbers.” And, although Descartes does not look at it this way, he would probably accept this response. However, he has another point in mind because he wants to capture the geometry by his use of algebraic expressions.

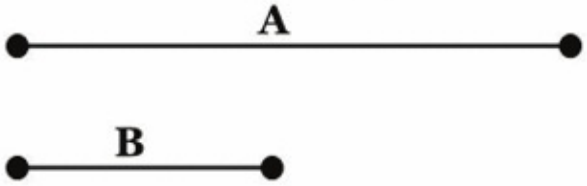
By the time the quoted statement appears, Descartes has already shown that, once a unit length has been introduced, one can, geometrically, multiply line segments and take square roots.[36]

And this is important because, in Greek geometry, one did not multiply line segments. Where a modern treatment invokes a product, Euclid would invoke a rectangle. And a square root was, implicitly, the length of the side of a square. However, Descartes manages to construct a line segment representing the product of two line segments and, in a separate construction, a line segment representing the square root of a given line segment. He seems to take his successes to imply that the results, the line segment he constructs in each case, embodies the same units as the original line segments.

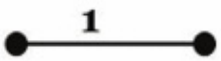
The underlying issues are the same in each of his constructions, so I will only examine the first of them here, namely Descartes’ construction of a product. My examination of his construction will check the validity of his claim. But it will also bring to light how his construction relates to my own analysis and why it fails to address the issue that I examined at the end of the last section.

Descartes’ multiplication of line segments proceeds in two steps as follows:

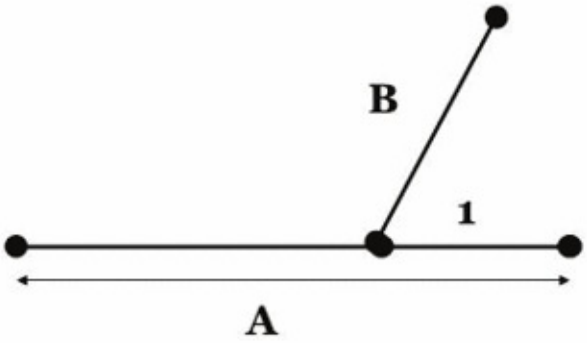
Segments to multiply

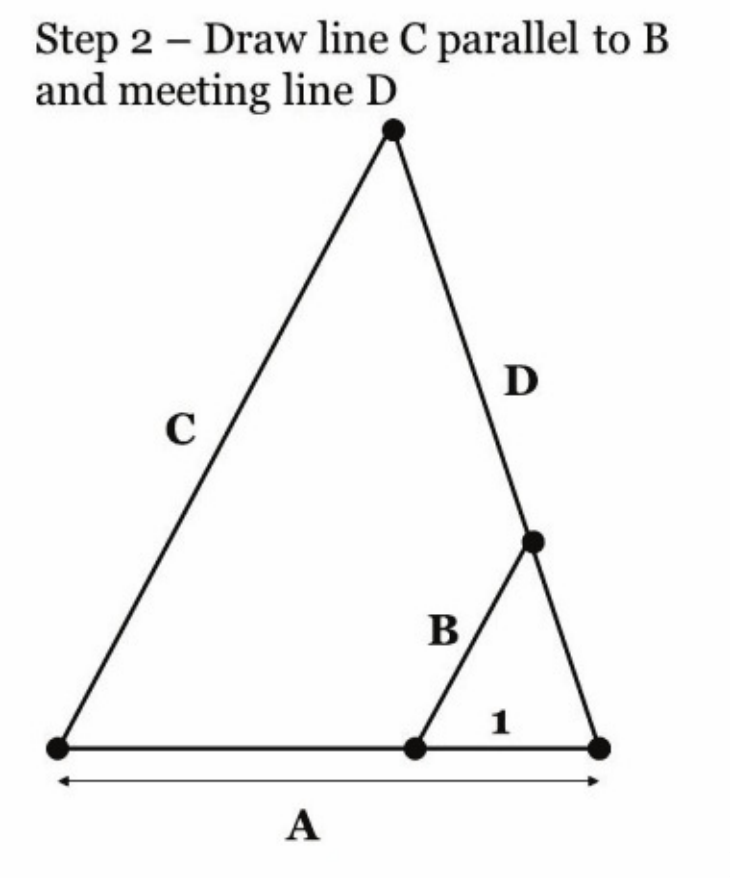


Unit segment



Step 1 – Lay the unit segment along A and attach B at an angle to the interior end of the unit segment.





By similar triangles, the ratios C/A = B/1.
If one now treats C, A, and B as numbers instead of magnitudes, one can reduce the equation to C = A × B/1 or C = AB.

But, in fact, A, B, and C are *not* numbers and Descartes' construction purports, indeed, to provide a multiplication of magnitudes, albeit a multiplication that depends on a choice of unit. So how should one look at this construction?

I will discuss Euclid's treatment of ratios such as B/1 in the next section, but the Greeks, at least implicitly, treated ratios as numbers. The ratio of magnitudes C/A is a number, though C and A be magnitudes. Indeed, that a ratio of magnitudes of the same kind is a number is implicit in all measurement. Were this not the case, one could not numerically relate a magnitude to another standard magnitude to express its measurement as a number. Thus, for example, the ratio B/1 is the *numerical* length of B as expressed in terms of the unit segment designated as '1'.

This recalls my earlier definition of ratio. In general, if X and Y are magnitudes and b is a number, then the formulas X = bY and X/Y = b can be regarded as two different expressions of the same relationship between X and Y. For example

different expressions of the same relationship between X and Y. For example X/Y = 3/5 identifies the same relationship as X = (3/5) × Y.

It is in this sense that one can deduce, from C/A = B/1, that C = (B/1) × A. In this formula, so far, A and C are still magnitudes and the expression on the right should be interpreted as multiplication of A (a magnitude) by (B/1) (a number) to yield C (a magnitude).

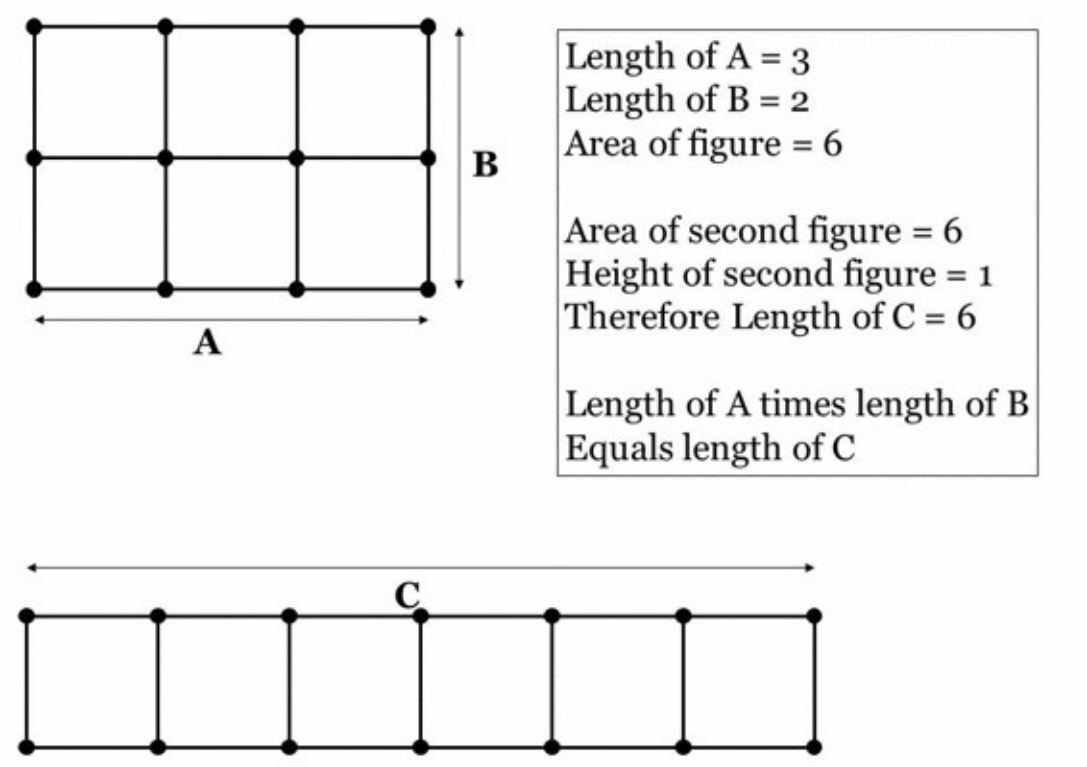
But the problem with this formula is that Descartes wants to treat A, B, and C on the same footing. For that, it is not enough to express B as it relates to 1 without doing the same for A and C. So taking the ratio of each side of the equation to 1, one arrives at

C/1 = (B/1) × (A/1) = (A/1) × (B/1)

This, at last, is a relationship of numbers and it is the *only* formula, following from Descartes' analysis, that also puts A, B, and C on the same footing. AB = C, to make sense, must be interpreted as a short hand expression for this formula and, insofar as Descartes might think he has done more than this, he would be equivocating (between numbers and measured magnitudes; between the *measure* of measured magnitudes and the *magnitudes* that one is measuring). Descartes has indeed produced a line segment. But he has *not* multiplied line segments, as such. What he has accomplished is to construct a line segment whose length, expressed *in terms of a given unit*, is equal to the product of the lengths of two other line segments, each expressed *in terms of the given unit*.

Properly interpreted, Descartes has exhibited a valid relationship even if his own interpretation of that relationship is unclear or open to question.

In the interests of better understanding that relationship, it may be helpful to tie it into my earlier analysis of area. So consider, again, segments A and B. Their product has the dimensions of area, an area that can be exhibited by creating a rectangle of sides A and B. Once a *unit* has been chosen one can create a second rectangle enclosing the same area as the first and for which one of the sides is the chosen unit. If, for example, A is 3 units and B is 2 units, the area of the first rectangle is 6 square units. Since one side of the second rectangle will have a length of 1 unit, the other side must be 6 units. The entire set of relationships is illustrated in the following diagram:



Descartes, then, has not addressed the abstraction of arithmetic from units. Rather, he has, in effect, adjusted the units by dividing one of the factors by, say, 1 foot to reduce an answer in square feet to an answer in feet – a step that is only possible when one has chosen a unit (in this case, 1 foot). That Descartes understands this much is clear. For it is made explicit, as his second major point, in the continuation of the passage that I quoted earlier. Descartes writes:

"It is not, however, the same thing when unity is determined, because unit can always be understood, even when there are too many or too few dimensions; thus, if it be required to extract the cube root of $a^2b^2 - b$, we must consider the quantity a^2b^2 divided once by unity, and the quantity b multiplied twice by unity."[37]

Descartes's point is that dimensions are not a problem as long as a choice of unit is always assumed to be given and all required adjustments are made at every turn, an assumption he makes thereafter, implicitly or explicitly, whenever he needs to make it. Having disposed of the issue in the first few paragraphs of his *Geometrie*, he never raises it again.

Descartes's goal was to reduce geometry to algebra. The price he willingly paid

to make the transition was to assume, thenceforth, that a unit has been chosen. He was right that selecting a unit was the price and it remains the price to this very day. It is the price we rightly pay whenever we choose coordinates and begin to calculate. Descartes accomplished what he set out to do insofar as he did succeed in finding an algebraic formulation of problems in geometry, an achievement of epic importance. But, although his concerns were related to mine, he was not addressing quite the same issue.

It is worth pausing a little longer to see how Descartes' coordinate approach plays out in elementary physics.

Consider the motion of a projectile, starting from ground level with an initial upward velocity. Restricting the discussion to the height x of the projectile above the ground and letting t denote time, the motion obeys an equation of the form $x = at^2 + bt$, where a is negative and b is positive. This is the equation of an inverted parabola and it remains valid (ignoring air resistance, etc.) until the projectile ultimately strikes the ground. The motion starts when $t = 0$.

But what about the diverse types of magnitudes involved in this equation? First, 't' represents time which, to begin with, does not have the units of distance. And here we are squaring the time dimension and then adding the squared value to time! What is going on?

Well, the key lies in the innocent coefficients 'a' and 'b'. The missing units are contained in 'a' and 'b'.

So what are these coefficients 'a' and 'b'? To begin with, they are not numbers and, to express them numerically, one needs to choose units. In general, to relate this equation, or any equation in physics to the world, one must invoke a set of standard units of measure, a set that provides for every term in the equation. So suppose one chooses feet for x and seconds for time (t). Having now chosen units of distance and time, and only now, one can identify the values of 'a' and 'b'.

The first coefficient, a, is a constant on the earth's surface. Its value to two significant figures is -32 feet per second per second and its full expression requires both the numerical value and the units in which it is expressed. This is the downward acceleration due to gravity; it means that, in every second the *upward* velocity decreases by 32 feet per second or, equivalently, the *downward* velocity increases by 32 feet per second. Notice that there are two 'per seconds' included in the constant 'a' to cancel the 'square seconds' represented by t^2. The entire term, at^2, has units of feet.

What about the second constant 'b'? Well, that turns out to be the initial upward

speed (at time t = 0) and, accordingly, has units of feet per second. Once again, the ‘per second’ in b cancels the ‘seconds’ in t. Consequently, the term, bt, also has units of feet. Once this is all understood and retained as a context, writing out these units every time becomes pedantic and gets in the way of the algebra. But it is worth doing so at least once. So, choosing 128 feet per second as the initial upward speed:

x feet = - (32 feet per second per second) × t^2 square seconds + (128 feet per second) × (t seconds)

As this relates to my earlier analysis, the lesson is clear. In any application to concretes or even to general physical settings, units must be chosen and the units must be consistent throughout. As this example illustrates, choosing consistent units is not an academic exercise. Although the *form* of the formula was not affected by the choice of units, the *coefficients* were. Both coefficients would have been different numbers had the equation been written, for example, in miles per hour.

This example illustrates how the connection is maintained when one applies coordinate geometry to physical problems. Descartes’ coordinate approach provides a way to convert problems in geometry and physics to algebra (and, ultimately, to differential equations). By taking units for granted, in the way Descartes suggested, one can simply ignore the units while one solves the mathematical problem. Yet the units are always present in the background and must be made explicit again whenever one tries to connect the mathematical result to the original problem in geometry or physics.

This example shows the difference between the algebraic viewpoint and the viewpoint of the physicist. More generally, understanding the way that the formulation of physical laws depends upon the choice of units (ultimately, of coordinate systems) has proven enormously helpful in the discovery of new physical laws. So, to better understand the connection between the mathematics and the physics, let’s convert the units in this equation to miles per hour.

First,

128 feet per second = (128 feet per second) × (1 mile per 5280 feet) × (3600 seconds per hour) =
87.3 miles per hour

Secondly,

32 feet per second per second = (32 feet per second per second) × (1 mile per 5280 feet) × (3600 seconds per hour) × (3600 seconds per hour) = 78,500 miles per hour per hour

I have rounded the acceleration, because 32 feet per second per second is only accurate to two significant figures.
The formula for the trajectory in miles (for x) and hours (for t) becomes:

x miles = - (78,500 miles per hour per hour) × t^2 square hours + (87.3 miles per hour) × (t hours)
Ignoring the units in the expression, but keeping them in mind, one writes
$x = -78{,}500t^2 + 87.3t$

Now, to an algebraist, this equation and the previous one represent two different algebraic expressions. If one uses them, say, to find the maximum value taken by x, the highest point reached by the projectile, one solves two different equations. But for a physicist, these equations are simply two different expressions of the same physical law, applied to the same physical situation. My interest here is to appreciate both perspectives, to understand the relationship between the measurements and the measurement standards, the systems of coordinates that they invoke. More generally, this is one of the tasks of a geometer, to study the relationship between these two perspectives on a more abstract level, as it applies to situations more complex, more interesting, and more far-reaching than this example.

This example is a somewhat pedantic beginning on one of the paths that leads to a subject of enormous consequence, the study of symmetries and invariants that has proven so valuable to both physicists and mathematicians. I say more about the measurement of symmetry in Chapter 8.

The dependence of the coefficients on the choice of units is a commonplace today. Descartes laid the groundwork, but he did not address the issue to this level. Certainly, he invoked a device to make the units consistent, but that’s where his interest died. The discovery of how to apply these expressions in physical problems was left to others.

Returning to this theme, what subtlety does one lose when thinking purely in terms of numbers? If one does not work to remember it, one loses the geometry.

One forgets the connection to actual quantity, because geometry is that link.

Arithmetically, one can legitimately perform any operation on two numbers and get another number. But the example of the projectile illustrates that whenever one applies these operations to a concrete problem, matters immediately get more complicated.

For example, I can only add or subtract quantities when they have a common unit. I cannot add apples and oranges, nor can I add velocity to a time interval. On the other hand, although I can *multiply* velocity times a time interval, the need for consistency arises in another form. I have no choice regarding the units of the resulting product.

If one multiplies speed in miles per hour times some number of hours, one always gets a distance in miles. 30 miles per hour times 2 hours is 60 miles. One could, of course, express the answer in feet, but that would require another multiplication: 60 miles times 5280 feet per mile equals 316,800 feet. The arithmetic in the first equation, 30 times 2 = 60, *only* applies if one expresses that result in miles.

Multiplying force times distance (a measure of work) illustrates the general way one keeps track. If one applies a force of 5 pounds to move something 8 feet, the amount of work is said to be 40 *foot-pounds*. One keeps track simply by repeating each unit in the final expression. What makes this practice specifically significant is the physical fact, which had to be defined and established, that applying a force of 10 pounds to move something 4 feet involved exactly the same amount of *work* as the application of 5 pounds to move something 8 feet. One needs to have established that the amount of work is the same in both cases precisely in virtue of the fact that the multiplications, 5 times 8 and 10 times 4, are the same.

It is easy to take arithmetic for granted, multiplication in particular. Yet its complexity has always been there because the world that it is used to measure is complicated. The relationship of numbers to the world has been mysterious, in part, because the relationship of concepts to reality, in general, is still a mystery to almost everyone. Taking a step back to examine the things we take for granted has always been difficult. It is difficult because one must, in the very examination, find a place to which one can take that backward step.

Mathematics is referential. To understand mathematics is to understand how it

Mathematics is referential. To understand mathematics is to understand how it relates to the world. And that relationship of mathematics to the world becomes more subtle the further one proceeds to greater and greater levels of mathematical abstraction.

The place to start is at the beginning, with geometry and arithmetic, carefully tracing the relationships of key concepts to the world that those concepts help us apprehend. We take our first big step when we realize that things are not as simple as they first appeared or were made to appear. Because we need to confront our confusions before we can begin to address them. The breakthrough comes when, having glimpsed the complexity, we begin to grasp that there really is a way to make sense of it all.

The Greeks can be an inspiration on both counts. But to really understand mathematical concepts, to understand them as capturing aspects of the world, one needs a better understanding of concepts in general. One needs an third alternative between the Platonist view that conceptual objectivity requires a separate world of mathematics versus the nominalist view that mathematical concepts are not inherently referential, but are a free creation of the human mind. One needs, in my view, the perspective provided by Ayn Rand's theory of concepts. And I point, as a manifestation of this need, to the fact that neither mathematicians nor philosophers have, in modern times, been able to adequately account for the relationship of mathematics to the world. Indeed, to all appearances, they have abandoned the challenge, equating any objective referential character of mathematics to Platonism.

Finally, for future mathematicians, there is a third step. There is value in following the standard treatments to understand just how, for example, the "distributive law", established first for positive integers, applies to negative numbers, fractions, and irrational numbers, as well. But this value requires that one first fully grasp, and never forget, the relationship of these concepts to reality.

Ratios in Euclid

A ratio is a measure of relative magnitudes. Ratios are the foundation of the measurement of magnitudes. It is because we can identify the ratio of the height of a man to a foot-long ruler that we can identify his height as six feet.

Euclid's Elements speaks of the ratio of magnitude X to magnitude Y and provides criteria to compare ratios between different pairs of magnitudes.[38] Ratios may be equal (in proportion) or one ratio may be greater than another one. Euclid's definition of equality and inequality of ratios covers cases in which the first pair of magnitudes is different in type from the second pair of magnitudes. Thus, one can say that a ratio of line segments is greater than, less than, or equal to a ratio of areas.

Although Euclid's formal treatment in Book V provides a very precise criterion for equality and inequality of *ratios of magnitudes* there is nothing in that Book to justify relating a ratio of magnitudes to a number, or even relating it to a *ratio* of numbers. Yet such a relationship seems to have been taken for granted among Greek mathematicians, both before and after Euclid, in their various rational approximations to the irrational numbers π and √2, both of them ratios of magnitudes. And, finally, in Book X, Proposition 5,[39] Euclid states, but does not adequately prove, that "Commensurable magnitudes have to one another the ratio which a number does to a number."[40]

My viewpoint diverges from Euclid's, drawing on my earlier discussion of the prearithmetic of magnitudes to identify the relationship between ratios and numbers. In the section entitled "The Axiom of Archimedes", I applied the Axiom of Archimedes to argue that, if X and Y are magnitudes of the same kind, then there is a number A such that X = AY to any required level of precision. In relation to this formula, I have defined A as the ratio of X to Y. And, when convenient, I express this relationship as:

$$X/Y = A$$

This notation is both appropriately suggestive and compact, so long as one keeps in mind that, while A is a number (specified up to materiality with respect to the product AY), X and Y are not numbers. To make this fully concrete, X and Y might both be the length of objects, the weights of objects, or the frequencies of two vibrating strings, all considered without regard to the specific units to which they might be related. Looked at in this way, X/Y = A is just another way to express the relationship X = AY.

Now, how does this relate to Euclid? I claim that it is entirely consistent with his definitions in Book V and provides a way to relate Euclid's formulation to a modern perspective. The key definitions are definitions 3, 4, 5, and 7, so we consider each, in turn. Definition 3 reads:

consider each, in turn. Definition 3 reads:

"3. A *ratio* is a sort of relation in respect of size between two magnitudes of the same kind."[41]
Manifestly,

1. X = AY (where A is a number) is a relation of two magnitudes in respect to size
2. A relation of this type specifically applies when, and only when, X and Y are of the same type.

Definition 4 reads:

"4. Magnitudes are said to *have a ratio* to one another which are capable, when multiplied, of exceeding one another."[42]

In the context of Definition 3, this is a weak form of the Axiom of Archimedes, applied, in general, to magnitudes of a particular type, which I have already discussed. It is a weak form insofar as it leaves open a possibility that Aristotle did not leave open: that two magnitudes of the same kind might not have a ratio. Euclid needs to use this property, but does not commit himself as to its universal applicability. On the other hand, Euclid implicitly recognizes a further point. I say this because Definition 4 is only applicable to pairs of magnitudes of the same type. One cannot say, for example, that any multiple of a length can either exceed or fall short of an area. Areas and lengths are two different types of magnitude and cannot be compared in this way. Euclid, clearly, was aware of this issue and it was important to him.

Definition 5 reads:

"5. Magnitudes are said to be in the same ratio, the first to the second and the third to the fourth, when, if any equimultiples whatever be taken of the first and third, and any equimultiples whatever of the second and fourth, the former equimultiples alike exceed, are alike equal to, or alike fall short of, the latter equimultiples respectively, taken in corresponding order."[43]

This one is rather a mouthful and requires unpacking. Suppose given magnitudes W, X, Y, and Z. Euclid offers a criterion that the ratio of W to X is the same as the ratio of Y to Z. In my notation, this equality of ratios is written W/X = Y/Z. This, in turn, is another way of saying that if W = AX and Y = BZ then A = B.

The problem Euclid is addressing with his formulation is that W and X may be "Incommensurate" in the Greek sense, meaning that the unknown, *A*, is an irrational number. Euclid cannot reliably find specific whole numbers, forming an exact fraction, to measure the ratio. So he is left, as are we, with finding a less direct way. His approach, one discovered by Eudoxus and exploited, in a somewhat different form, by Dedekind in the 19th century,[44] is to locate that ratio between all the ratios of whole numbers that exceed it and all the ratios of whole numbers that are smaller. The awkwardness of Euclid's expression is due to a dearth of means to say this easily. Euclid's approach will become clearer as I unpack his formulation and relate it to my own definition of ratio.

To begin, following Euclid's definitions for equality of ratio, assume that W and Y are multiplied by m and X and Z are multiplied by n. There are three cases:

1. mW > nX, in which case Euclid's criterion says that
mY > nZ
2. mW = nX, in which case Euclid's criterion says that
mY = nZ
3. mW < nX, in which case Euclid's criterion says that
mY < nZ

Equivalently, these alternatives are

1. W > (n/m)X, in which case Euclid's criterion says that Y > (n/m)Z
2. W = (n/m)X, in which case Euclid's criterion says that Y = (n/m)Z
3. W < (n/m)X, in which case Euclid's criterion says that Y < (n/m)Z

Since W = AX and Y = BZ, each case can be written, in turn, as

1. AX > (n/m)X, in which case Euclid's criterion says that BZ > (n/m)Z
2. AX = (n/m)X, in which case Euclid's criterion says that BZ = (n/m)Z
3. AX < (n/m)X, in which case Euclid's criterion says that BZ < (n/m)Z

From which it follows, for each case in turn, that

1. A > (n/m), in which case Euclid's criterion says that B > (n/m)
2. A = (n/m), in which case Euclid's criterion says that B = (n/m)
3. A < (n/m), in which case Euclid's criterion says that B < (n/m)

In sum, *A* is greater than n/m only if *B* is; it is equal to n/m only if *B* is; and it is less than n/m only if *B* is. But m and n are chosen freely. So n/m could be any

positive rational number. Therefore, it follows from these alternatives that there is no rational number between A and B. But according to the Axiom of Archimedes, were A and B not equal, there would have to be a rational number between them. Since there isn't, then A = B.

Euclid's criterion, as it applies to my formulation of ratio, implies that the two ratios are the same because it implies that the numbers that measure them, A and B are equal. The Axiom of Archimedes is an essential underpinning of this argument.

Euclid's Definition 7 is simply a variation of Definition 5. It states:

"7. When, of the equimultiples, the multiple of the first magnitude exceeds the multiple of the second, but the multiple of the third does not exceed the multiple of the fourth, then the first is said *to have* *a greater ratio* to the second than the third has to the fourth."[45]

Not to belabor the point, in terms of the notation used to discuss Definition 5, this says that, for some values of n and m,
A > n/m but B < n/m

So the rational number n/m lies between A and B. It follows that A > B, which is to say that W/X > Y/Z.
This discussion has shown that my discussion of the prearithmetic of magnitudes, drawing, as it does, on modern discoveries relating to irrational numbers, suffices to define ratios. My definition is also consistent with the use of ratios in Greek practice. But, unlike Euclid's discussion, the modern perspective makes explicit, from the very outset, the relationship of ratios to numbers.
By doing so, my definition clears up one of the mysteries in Euclid's presentation. By the end of Book V, Euclid has demonstrated the basic properties of ratios. And Book VI proceeds to use ratios to prove geometric theorems *including*, *especially*, theorems relating ratios of areas to ratios of magnitudes. What gives him that right?
Well, the technical answer is that Euclid's definition of equal ratio in Book V in no way restricts him from relating ratios of different kinds of magnitudes. But, in the context of Euclid's Book, his ability to do so must seem like an accident. My perspective dispels this impression. In the equation W = AX, where W and X are magnitudes of the same kind, the quantity A is a number. But it's also

what physicists call a dimensionless quantity. It's a number of repetitions, first; a number of divisions, second; or a combination of the two (a rational number), third. Or, finally, it may be an irrational number that can be approximated, to any meaningfully required accuracy, by a rational number.

These basic operations, repetition and division, are the same, and have a corresponding result, no matter what kind of magnitude or number they are applied to. When a magnitude is multiplied by such a number, the result is a magnitude of the same kind.

In effect, the quantity A *measures* the relationship, the *proportion*, between the magnitudes. If A = 2, it means that W is twice X. If A = 1/3, it means that W is one third of X, regardless of the units of W and X. The quantity A, then, means exactly the same thing, specifies the same relationship or proportion, no matter what specific magnitudes are being related. So, *of course,* it can be used to relate ratios of different kinds of magnitudes.

Finally, my definition makes the various properties of ratios easier to prove and easier to understand.

Take an example from Euclid's Book V. Now most of the early propositions in Book V, which lay the groundwork for the more difficult ones, are immediately obvious from my perspective. But Proposition 16 is not obvious and it is also one of the most important.

Proposition 16 is applicable whenever W, X, Y, and Z are *all* magnitudes of the same kind (a qualification Euclid omits, presumably as an oversight). According to the proposition,

If W/X = Y/Z, then W/Y = X/Z

As Euclid puts it:[46]

"If four magnitudes be proportional they will also be proportional alternately."

A modern proof is algebraic. Suppose that W/X = Y/Z = A or, equivalently, W = AX and Y = AZ. Since W and Y are magnitudes of the same kind, there is a number B that expresses their proportion. Namely,

W = BY and, equivalently, W/Y = B It follows that AX = W = BY = BAZ, from which AX = ABZ and, therefore, X = BZ or, equivalently, X/Z = B

So W/Y = B = X/Z, proving the proposition.

In the context of Greek mathematics, the theory of ratios plays a central role. It is the form in which they drew comparisons among geometric magnitudes and confronted the issue of measurability. But their theory was never completely

integrated.
Thus, Euclid's Books V and VII present two different theories of ratio, one for magnitudes and one for numbers (meaning positive whole numbers). A ratio between magnitudes seems very like a number, especially when Euclid proceeds to compare ratios of lengths to ratios of areas and even, in Book VII, uses line segments as illustrations standing for numbers. Yet this connection is not made explicit until Book X and then only for ratios of commensurate magnitudes, and also without a complete proof.[47] It is not a complete proof because Euclid does not bridge the gap between the two different definitions of ratio.
Book V is particularly frustrating because its definition of equal ratio is a creation of genius. But the discussion is left hanging in mid air because Euclid is unable to state, lacked the vocabulary to state, that a ratio is a number. He tells us when ratios are equal, but he does not tell us what they *are*.
From a modern perspective, most of this mystery is simply unnecessary. The Greeks certainly knew how to multiply and divide line segments by numbers; much of their thinking was in those terms and they use those very concepts to define equality of ratios. From today's vantage point, which admits fractions and irrational numbers as bona fide *numbers*, only a very slight further development is needed to incorporate the entire subject of geometric ratios into what I have termed the prearithmetic of magnitudes. And my development of this prearithmetic of magnitudes was directly inspired by Euclid, together with Aristotle's fromulation of the Axiom of Archimedes.
But the Greeks did not have the benefit of our perspective. Qua numbers, they had no concept of irrational numbers, no systematic vocabulary to specify them. And, for them, irrational numbers always arose in geometric contexts, as the incommensurate relationship of two magnitudes, such as the relationship of a diagonal to the side of a square. Geometric ratios, in Euclid, were specified, always, by a pair of magnitudes, usually lengths.
The Greeks could, indeed, specify a ratio of commensurate magnitudes by a pair of numbers counting their respective common units. So, for example, lengths of 2 inches and 7 inches bear a ratio of 2 to 7 counting, respectively, their common unit, namely inches. But the Greeks had no systematic numerical way to designate irrational numbers, no vocabulary for doing so.
They were left to address the fundamental problem of measuring incommensurate relationships and they started with the object of such measurement: pairs of incommensurate magnitudes. And they answered the first, all-important question: When are two pairs of magnitudes in the same ratio? Or, alternatively, when is one ratio greater than the other? And their answers, arguably, were eventually the inspiration for Dedekind's account of irrational

numbers in the 19th century.

[1] Neal H. McCoy, *Introduction to Modern Algebra*, Boston, Allyn and Bacon, Inc., 1960, chapters 4-6, for the usual development

[2] Carl B. Boyer, *History of Analytic Geometry*, Mineola, NY, Dover Publications, Inc., 2004, especially p 74-102, regarding Descartes and Fermat as the inventors of analytic geometry

[3] Euclid, *Elements*, edited with notes by Thomas L. Heath (New York: Dover Publications, 1956), especially Book V. Also, Archimedes, *The Works of Archimedes*, 1897, Cambridge: at the University Press, Books 1 and 2 of "On Equilibrium of Planes" and Books 1 and 2 of "On Floating Bodies"

[4] Ayn Rand, Introduction to Objectivist Epistemology Expanded Second Edition, p 7 in the paperback edition

[5] Leave aside a) physical underpinnings b) whether it's ultimately reducible to a magnitude

[6] Rand, p 12, on omitted measurements

[7] The theory of real numbers studies the arithmetic and other relationships of rational numbers, irrational numbers, and negative numbers

[8] Euclid, Book V

[9] Euclid, Book V, Definition 4

[10] Archimedes, "On Equilibrium of Planes 1" Propositions 6 and 7

[11] Euclid, Book I, Proposition 47, "In right-angled triangles the square on the side subtending the right angle is equal to the squares on the sides containing the right angle." Euclid proceeds to draw the squares, draw a number of auxiliary lines (an abstract form of measurement) and, from that, argue to the conclusion

[12] Sir Thomas Heath, *A Manual of Greek Mathematics*, Dover Publications 1963, Ch VII, p 150, in a section entitled "The Conchoids of Nicomedes", The conchoid is a curve, constructed by a mechanical device (around 270 B.C.) for the specific purpose of solving this sort of problem

[13] Those with a mathematical background may notice that this entire setup is simply a way of treating magnitudes as a onedimensional vector space.

[14] Aristotle, *Physics*, Book III, Section 6, at 206^b1, lines 12 – 13, quoted from the Revised Oxford Translation, edited by Jonathan Barnes, Princeton University Press, 1984

[15] Heath, *A Manual*, See the discussion of "The Method of Exhaustion" in Chapter VIII, p 193

[16] Pat Corvini, "Achilles, the Tortoise, and the Objectivity of Mathematics", lecture series delivered at the Objectivist Summer Conference, July, 2005 and available on CD from the Ayn Rand Book Store at www.aynrandbookstore.com, Corvini begins by pointing out that, as applied to any concrete case, there is always some limit to how far one can go with these subdivisions. The division can continue until it is no longer possible to distinguish the subdivisions from the magnitude being divided. A mathematical concept always applies to each case insofar, and to the extent, that it is possible to make the relevant distinctions. However, the mathematical concept itself, like all concepts, is open-ended and applies universally, in just this way, to all cases. As Corvini would put it, the limits are determined on the side of each concrete instance, not on the side of the mathematics. It is precisely those physical limits that are treated as omitted measurements in the process of forming mathematical abstractions involving continuous quantities

[17] On the side of the mathematics, the Axiom of Archimedes implies that there are an unlimited number of numbers – whatever number of numbers one needs to measure any magnitude one encounters by a chosen standard. Yet its content, as applied to concretes, is precisely the reverse: It implies that all magnitudes are

measurable – that there can be no magnitude larger than any possible multiple of a chosen standard. There is only an apparent conflict, here, if one confounds the abstraction with it units and, thus, confounds mathematical infinity with an actual infinity. This is an example of a point Corvini made in her lectures, that the dangers of confounding abstractions and the unit of those abstractions is particularly dangerous in mathematics and, specifically, when thinking about mathematical infinity

[18] Rand, p 13. "The element of *similarity* is crucially involved in the formulation of every concept; similarity, in this context, is the relationship between two or more existents which posses the same characteristic(s), but in different measure or degree." Three sentences later: "All conceptual differentiations are made in terms of *commensurable characteristics* (i.e., characteristics possessing a common unit of measurement). No concept could be formed, for instance, by attempting to distinguish long objects from green objects." The discovery of the relationship of measurement to concept formation is one of the fascinating aspects of her theory

[19] To avoid confusion, one must distinguish the concept "commensurable", as used by Ayn Rand, from the Greek concept of "commensurate". The Greeks considered two magnitudes to be commensurate if their ratio was a rational number. Otherwise, they considered them incommensurate

[20] One will, nonetheless, find such proofs in the mathematical literature. But that only means that the axiom has been already introduced in some other form – as when Euclid includes it as Definition 4 in Book V

[21] Rand, chapter 6 on "Axiomatic Concepts"

[22] Rand, p 196. It is here that Ayn Rand's approach to precision is particularly striking. She says, 'But more than that, isn't there a very simple solution to the problem of accuracy? Which is this: Let us say that you cannot go into infinity, but in the finite you can always be absolutely precise simply by saying, for instance: "Its length is no less than one millimeter and no more than two millimeters."'

[23] Corvini

[24] Archimedes, "On Equilibrium of Planes 1," Propositions 6 and 7

[25] David Harriman, The Logical Leap: Induction in Physics (New York, New American Library, 2010), p 47-8 in Chapter 2, "Experimental Method"

[26] Euclid, propositions I.47 and II.13

[27] I discuss Euclid's analysis of area in Chapter 3

[28] Archimedes, "On Equilibrium of Planes," p 189-220 and "On Floating Bodies," p 253-300

[29] David Harriman, Chapter 4, especially the section "The Development of Dynamics," p 117-30

[30] Robert Resnick, and David Halliday, Physics, Part I, New York, John Wiley & Sons, Inc., 1966, p 83, section 5-3 "Force"

[31] Resnick and Halliday, p 193

[32] Rand, p 12-13, In proceeding to this level of abstraction, I appeal explicitly to Ayn Rand's theory of measurement omission. On page 12, she says, "Bear firmly in mind that the term "measurements omitted" does not mean, in this context, that measurements are regarded as non-existent; it means that *measurements exist, but are not specified*. That measurements *must* exist is an essential part of the process. The principle is: the relevant measurements must exist in some quantity, but may exist in any quantity." Then, on page 13, "A concept is a mental integration of two or more units possessing the same distinguishing characteristic(s) with their particular measurements omitted." In my application, the omitted measurement is "units" and no requirement is made that each number in the calculation has the same units. One requires only a consistency in the use of the units.

[33] Jeremy Gray, *Plato's Ghost: the Modernist Transformation of Mathematics*, 2008, Princeton, Princeton

University Press, Apparently, Newton held this view. On page 134, Gray refers, in passing, to the "Newtonian view that the real numbers were ratios of quantities."

[34] Rene Descartes, *Des matiers de la Geometrie*, 1637, available in English translation as *The Geometry of Rene Descartes*, Dover Publications, Inc., 1954

[35] Descartes, p 6

[36] Descartes, pictures and original text on page 4, English translation on page 5, entire discussion occupying the second, third, and fourth short paragraphs of his treatise

[37] Descartes, p 6

[38] Euclid, Book V

[39] Euclid, Book X, Proposition 5

[40] Euclid, Volume 3, p 25, as Heath points out in his notes, The main problem is that Euclid has provided a different definition of ratio for numbers, in Book VII than he does for magnitudes in Book V. Euclid is obliged to show how these definitions relate

[41] Euclid, Book V, Definition 3

[42] Euclid, Book V, Definition 4

[43] Euclid, Book V, Definition 5

[44] Richard Dedekind, *Essays on the Theory of Numbers*, Dover Publications 1963 from a 1901 English translation, German publication 1872

[45] Euclid, Book V, Definition 7

[46] Euclid, Proposition V.16

[47] See note 150

Chapter 3 Geometric Area, Proportion, and the Parallel Postulate

The Parallel Postulate

All of Euclid's postulates formulate perceptual observations. Properly understood, their fundamental importance relates to their measurement implications. Their application to the world, to actual problems of measurement, is, in every specific case, subject to contextual precision requirements. They apply universally to any concrete case for which the required precision is actually achievable.

Historically, however, the Parallel Postulate has received enormous scrutiny compared to the other postulates.
Prior to the discoveries in the early 19th century, this scrutiny involved persistent attempts to prove the parallel postulate from the other postulates.[1] Postulate Five received particular interest because it seemed to invoke infinity in a way that the others did not.
Less noticed was a similar invocation in Postulate One. If Postulate Five asserted that two *converging* lines would *ultimately* meet, Postulate One, by implication, by what Euclid took it to mean, asserted that two *diverging* lines would *never* meet.
Ultimately, the development of nonEuclidean geometries, challenging these attempts to prove Postulate Five, led to the conclusion that Postulate Five was, in fact, independent of the other four postulates. These geometries satisfied Postulates One through Four, but not Postulate Five. The relevance of this work gained acceptance, especially, with the development of geometric structures, residing in ordinary Euclidean space that were interpretable as nonEuclidean geometries, as satisfying the Postulates of these nonEuclidean geometries.[2]
One might counter that such examples amounted to a redefinition of the concept of a straight line. But the examples remained impressive. For the examples implied that nonEuclidean geometries, in which Postulate Five did not hold, were fully as consistent as Euclidean geometry, consistent because they could be subsumed under Euclidean Geometry.[3] One could, then, realize nonEuclidean geometries without contradiction. One could not, without appeal to some

premise or observation about lines, circles, and direction, beyond Euclid's first four postulates, prove Postulate Five.

Such examples had one important thing in common: On a sufficiently small scale, the lines, circles, and directions in these examples looked like Euclidean lines, circles and directions. The situation is similar to the measurements one makes on the surface of the earth. On a perceptual scale, the earth appears flat and measurements of the earth satisfy Euclid's axioms. But the earth is not flat and, on a larger scale, its curvature becomes important and must be included in one's calculations.

In the nineteenth century, these discoveries were enough to challenge Euclidean geometry in its historical role as the ultimate foundation of mathematics. Near the end of the previous century, with amazing timing, Kant had enshrined space as the form of perception, as *a prior*, as the "form of all appearance of outer sense."[4] But, unbeknownst to Kant, the scientific basis for any such claim had already collapsed.

Notwithstanding, Kant's world of phenomena easily survived the introduction of nonEuclidean geometry. The casualty was on the side of mathematics. As we shall explore futher in Chapters 4 and 6, the birth of nonEuclidean geometry set the stage for the struggles over the foundations of mathematics that were to follow. And the most important philosophical backdrop for that struggle was provided by the phenomenology of Kant.

In regards to nonEuclidean geometry, the twentieth century added a final twist with Einstein's development of the General Theory of Relativity. General Relativity is built on the geometry of light rays, a geometry in which light rays, propagated across space are treated as straight lines. And the geometry of these straight lines, according to Einstein's theory, is not Euclidean.[5]

From a realist, measurement-centric perspective, and aside from their historical interest and import, these developments are important to the validity of the parallel postulate only insofar as they *circumscribe* its applicability. They are part of the context that one must consider and they certainly have measurement implications. They bear on the *context* of the Parallel Postulate, and they also create a need to expand from that base and even point to the direction that such an expansion must take, but *they do not affect the validity of the Parallel Postulate within its proper context*. Euclidean geometry remains the geometry of the perceptual scale to which all of man's measurements ultimately relate.[6]

Context and the Parallel Postulate

On the perceptual level, one can see that the opposite edges of a rectangular table point in the same direction. And someone in the living room can look in the same direction as someone else in the kitchen. These are perceptual identifications, unambiguous within any standard of precision that might be required at the perceptual level.

But to compare the rotational axis of the earth with the rotational axis of Mars, requires more elaborate physical means. And, in general, when one goes beyond direct perceptual observation, one needs to take these physical means into account to establish a meaningful comparison. As Ayn Rand puts it, "When you speak of measurement, you always have to define contextually your method of measurement."[7]

On an astronomic scale, light rays (or electromagnetic waves) are an essential part of all geometric measurement. The path of a light ray through space is the straightest path known to man. As such, a light ray is used to establish a line of sight, a specification of the direction from earth to other objects in the universe.

But the path of a light ray is not always straight. For example, light rays bend, are refracted, when passing from one medium to another. More critically, light rays bend when passing through gravitational and other fields. To understand the meaning of a measurement by light rays requires understanding the effects of such factors. For example, in the case of refraction, this means understanding the angle of refraction.

The effect of gravitation is more subtle. We know that the direction of a light ray, as it approaches earth from a distant star, will depend, for example, upon the position of the sun in relation to the direction of the distant star. Our best current understanding of just how the position of the sun will influence this angle is provided by the General Theory of Relativity. So the relevant context in measuring the direction of a distant star involves, first of all, the fact that one is measuring the direction of the distant star as the light ray flies. Secondly, the context includes the position of the sun and of any other massive bodies between us and the distant star. Thirdly, it includes our understanding of how various massive bodies influence our measurements by means of light rays. Taken together, our determination of the direction of incoming light rays from the distant star is a specification, as such, of the direction of the distant star, the only kind of specification that is physically possible today. And if we had some alternative way of measuring its direction, that alternative would be just that, *an*

alternative way of measuring its direction that would involve a different set of contextual factors.

I mentioned as a third factor “our understanding of how various massive bodies influence our measurement.” That understanding, however, involves solving, at least approximately, the Einstein Field Equations, a nonlinear set of equations for which very few exact solutions are known. To what extent is that an issue? And how does it matter?

First, there is a special circumstance in which the effect of the sun can be pretty much eliminated. If the sun, at one instant is on one side of the star and, at the next, is on the other side, one can thereby quantify the influence of the sun and, by taking the average of the two positions, eliminate it. Indeed, this circumstance, exploiting a total eclipse of the sun, provided important validation of Einstein’s Field Equations. However, the gravitational field of the sun affects the observed angles of all incoming light rays, even rays that don’t pass so near the sun, and the Field Equations are needed generally to quantify the effect of the sun’s mass on all incoming light rays.

Secondly, specifying the direction of the light ray and of the factors that are known to influence the result of one’s measurement does , in fact, *specify* the direction of the distant star, so long as one maintains the *context* of that measurement. However, without the use of the Einstein Field equations to quantify the effect of the sun on incoming light rays, one has no way to integrate that knowledge. One can specify that direction today, but one would have no way of relating the measurements one makes today to the measurements one will make tomorrow when the sun is in a different position. So one’s understanding of the direction would be limited to a very special context, namely the particular position of the sun at the time of the measurement. It is Einstein’s Field Equations that provide the critical link; that integrates one’s measurements and makes them meaningful, taken as a whole.

With all that said, one’s measurements of the directions of the stars and their distances, based upon the light reaching us from those stars, offers a coordinate system applicable to the universe at large. Spatial relationships in the universe are measurable and the base of their measurement is Euclidean geometry.

These light-coordinates do not, however, provide a *Euclidean* coordinate system. For example, one cannot simply rely on the Pythagorean Theorem, together with

one's measurements of the direction and distance of various stars, to compute the distance between any two of those stars from each other. Not in isolation from the laws of physics! One needs relativistic corrections to make these determinations.

So the question presents itself: Is there some formula, some recipe for appropriately adjusting our measurements of distances and direction? Is there a universal way to convert our measurements of stellar position to the measurements that a Euclidean coordinate system would provide? Is there a way to compute the "real" unadulterated "measurements" of direction and distance?

Now, obviously, such a recipe would be a sort of relativistic correction; a way of identifying the impact of massive bodies on the paths of light rays. Such a correction would be similar to the way one adjusts for refraction of light on earth. Having made these corrections, if these were possible, one would be in a position to apply the Pythagorean Theorem (in its three dimensional version) to find the distances between any two stars in the galaxy. And one could apply standard trigonometry to find the angles in any triangle formed by any three stars in the galaxy.

But there is a final requirement for such a formula: One's recipe would need to provide consistent answers regardless of which planet or star one took as one's vantage point. It would be illegitimate to say, for example, that a determination of distances and angles based on data acquired on Sirius would, *by the same recipe,* give different answers from a determination based on data acquired on earth. The distance between two stars cannot depend upon which *Euclidean* coordinates one uses.

Simply put: It's certainly not obvious that this can be done and it may very well be impossible. It's a problem that involves both physics and mathematics. Physically, it has to apply a physical theory such as General Relativity to account for the effect of gravitation on light. Then with, say, General Relativity as a given, one has to solve a mathematical problem that is, at best, non-trivial, and may be impossible. The important point is this: This is a scientific question, not a philosophical one and not fully even a mathematical one.

But one need not settle this question to establish the validity of one's measurements. Nor would its resolution alter our reliance on Euclidean geometry, including the parallel postulate, even as we make relativistic

corrections. Euclidean geometry is the frame of reference for *interpreting* relativistic corrections. Because Euclidean geometry is the geometry that applies on a *perceptual* scale, the scale to which all of our measurements must ultimately relate.[8]

And what about Euclid's criterion for parallel lines, the criterion that, in reference to Figure 1, the angles 1, 2, and 3 are equal?

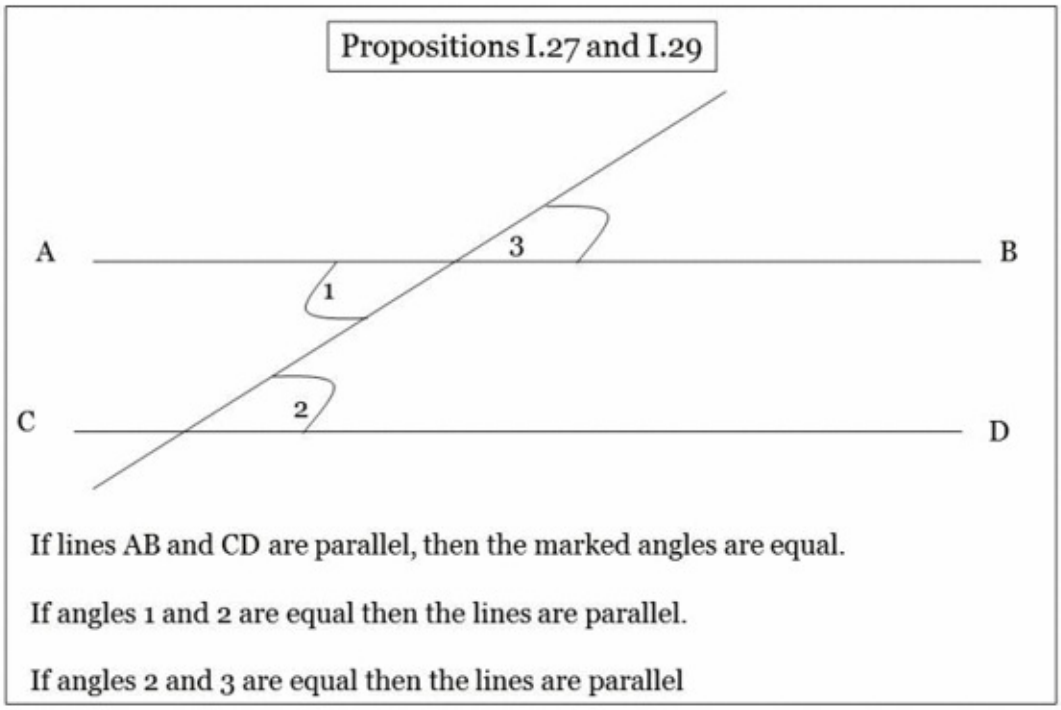


Figure 1

As it happens, in his proof of Proposition 27, Euclid proceeds by contradiction: If the two lines meet within the relevant threshold then angle 2 cannot equal to either angle 1 or angle 3. Euclid supposes that two lines meeting his criterion were to intersect and then argues that this would contradict the criterion. Euclid's demonstration is valid, just as his demonstrations of earlier propositions, provided it is taken contextually. It is valid on any scale and precision level on which a) one actually achieves the required precision level and b) his postulates are applicable to the means by which the relevant measurements are made.

Why is Postulate 5 Independent from the other Postulates?

In light of its measurement implications (and leaving aside the verdict of history), one should expect Postulate 5 to be independent of the other postulates. Postulate 5 states a condition for two straight lines to intersect. It says, in effect, that converging lines in the plane will continue to converge and ultimately intersect, manifesting their difference in direction. Postulate 5 provides a criterion for a very restricted judgment, a judgment that two lines are pointing in different directions. But of Euclid's postulates, only Postulate 5 provides any basis whatsoever to compare directions from differing vantage points.

Figure 2 illustrates the independence of Postulate 5 from the other Postulates. It addresses an attempt to define direction globally, by starting with a reference line as a standard direction. In this way, one might simply *define* parallelism, without the benefit of Postulate 5.

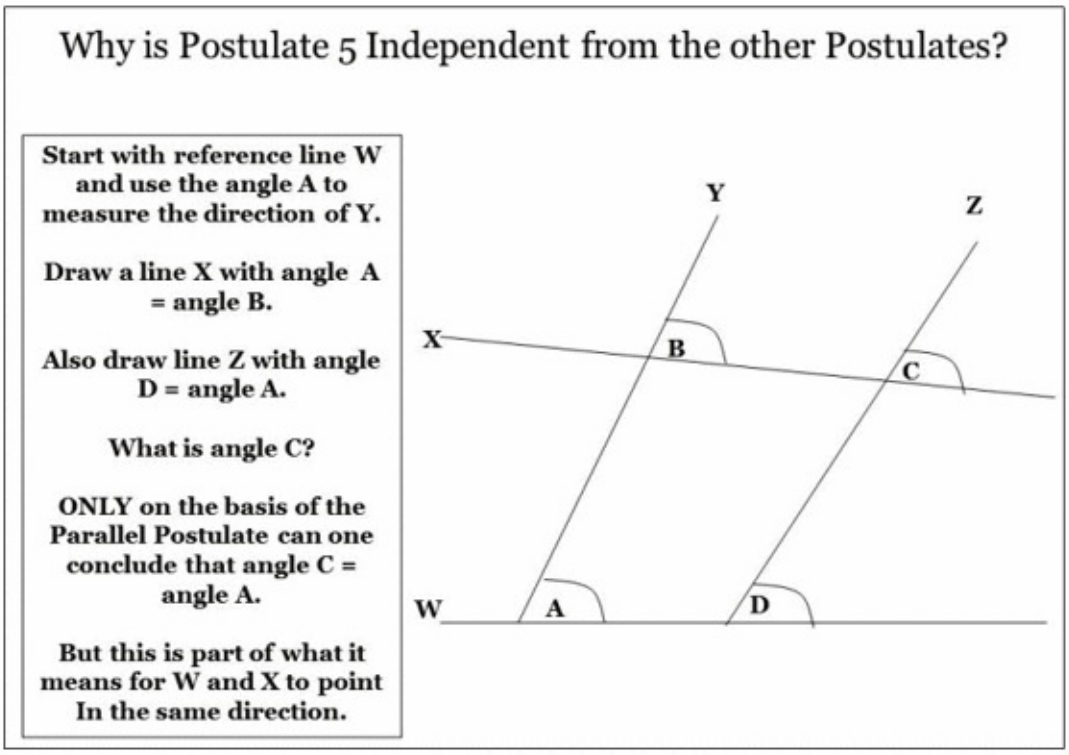


Figure 2

Had I drawn this accurately, it would show two intersecting pairs of parallel lines. However, I have deliberately distorted the drawing to make it necessary to focus *conceptually* on the relationships involved.

Euclid's Propositions 27 and 29 establish that a straight line cutting two other lines will make the same angle with both if and only if those two lines are parallel. OK, suppose that one starts with this circumstance and tries to use it as

a definition, without the aid of the specific warrant of Postulate 5 or some other fifth postulate. Suppose one simply says that two lines are parallel precisely when their intersections, with some chosen reference line, cut the same angle with the reference line.

At first glance this approach might seem enough to define direction, unambiguously, everywhere. But does it? Absent Postulate 5 and the intended implications of Postulate 1, does one really have a warrant for thinking so? On one condition only: *if* it turns out that this criterion, that two lines point in the same direction, be *independent* of the choice of reference line.

The postulates are silent on this. However, Euclid does provide such a warrant, namely Propositions 27 and 29, taken together. But these propositions depend, as least as far as Euclid's arguments are concerned, on his five postulates and on Postulate 5 in particular. Euclid proves Proposition 27, on the basis of his interpretation of Postulate 1 that there is a *unique* straight line connecting any two points.[9] But the importance and full meaning of the concept of parallel lines depends on the converse of Proposition 27, namely Proposition 29, an elaboration of Postulate 5. Parallel lines (i.e., lines in a plane that never meet) matter, they represent a common direction, not just because they *exist*, but because that existence is *unique*.

So what goes wrong if one is not allowed to use these propositions? For easier reference, I repeat Figure 2:

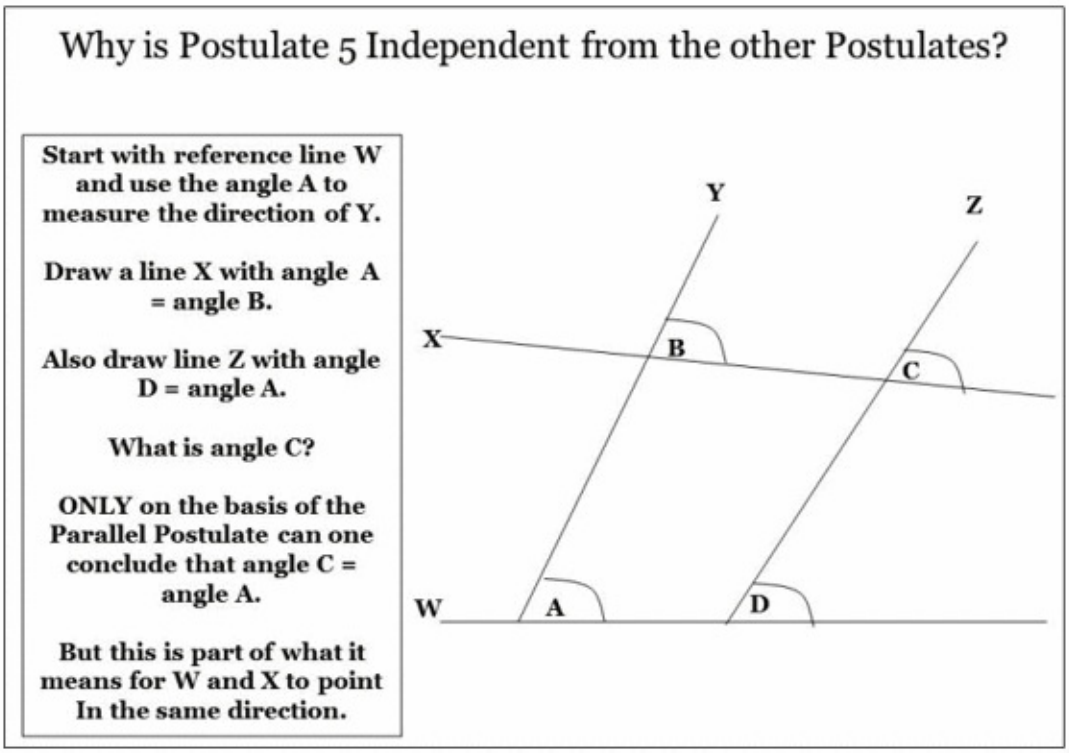


Figure 2

Let line W, in Figure 2, be the chosen reference line. Suppose that lines Y and Z make the same angle with W, that angle A = angle D. Now choose a different reference line X. For the sake of the argument, select a line X parallel to the reference line. Which means selecting a line X such that angle A = angle B.

But what is angle C? Proposition 27 implies that lines X and W will never meet. But it does not imply that angle C = angle B or that angle C = angle D: That would require Proposition 29. But Proposition 29 depends on Postulate 5; the first four postulates are not enough. There is no other known basis, deriving from Euclid's first four postulates, to claim that angle C = angle B.

There is simply no way, except by virtue of the Parallel Postulate or the equivalent, that one can *guarantee* that the angle C is equal to the other three angles. With respect to the reference line W, the lines Y and Z are pointing in the same direction. To demonstrate that Y and Z are also pointing in the same direction with respect to a *different* reference line such as X, would require more than the first four postulates.[10] Granted, by Proposition 27, derived from the first four Postulates, the lines will never meet. But, without Postulate 5, how does that establish that angle C is equal to the other three angles?

An unambiguous global determination of direction depends upon the Parallel

Postulate. One must have already made a perceptual identification such as Euclid's fifth postulate. And one must have proven Propositions 27 and 29 to validate this method of measuring out parallel lines.

Attempts to prove the Parallel postulate from the other postulates has a long history. That history ended in the nineteenth century with several independent discoveries that such a proof is not possible.[11]

There is a kind of surface known as a hyperbolic surface. Roughly speaking, from an external perspective, it looks like a saddle at every point. But, on a sufficiently small scale, it looks flat. The first four Euclidean Postulates all hold on a hyperbolic surface, without qualification, and they mean essentially the same thing that they do in Euclidean geometry. In particular, they have the same measurement implications.

In Chapter 1, I discussed geodesics in regards to the earth as they relate to Postulate 1. A geodesic is a line that, on a small enough scale, does not curve or bend; that looks straight, that keeps going in the same direction along the surface. Just as great circles serve as geodesics on the earth; just as the lines that we draw on a concrete slab on the earth do not curve or bend and look straight; a hyperbolic surface has geodesics. Any two points on a hyperbolic surface can be connected by a unique geodesic. Any geodesic can be extended, as needed, in either direction. A circle of prescribed radius can be drawn at any point on the hyperbolic surface. And all right angles are equal.

Yet Postulate 5 fails on this surface. It is entirely possible on a hyperbolic surface for geodesics that are approaching each other at one point to ultimately veer off without ever intersecting in either direction. So, as a matter of deductive logic and of measurement, Postulate 5 must be independent of the others. So far as Euclid's postulates and the interpretation of those postulates are concerned, a hyperbolic surface differs from a flat plane only in that Postulate 5 is valid on the plane and invalid on a hyperbolic surface.[12]

The Geometry of the Earth

The word "geometry" derives from the Greek words *ge*, meaning "earth", and *metria*, meaning "to measure". It was used by the ancient Egyptians to do just that: in order to reestablish property boundaries after each flooding of the Nile

River.[13] Geometry owes its name to one of its earliest uses: measurement of the earth.

The earth is a curved surface, yet on a small scale it looks as though it were flat. And, to add to the confusion, we have established a coordinate system on the Earth, a nonEuclidean coordinate system, consisting of longitudinal lines running North and South and latitudinal lines running East and West. At any point, except for the two poles, the latitudinal East-West lines intersect the longitudinal North-South lines at right angles.

This coordinate system on the earth is radically different than the coordinate systems one draws on a sheet of graph paper. Yet it is easy to forget the distortions in our flat maps of the world. These maps employ a “Mercator” projection of the globe onto the plane, a projection that preserves angles, but distorts lengths. The map will accurately show that one city is southwest of another city, but it will distort the distance between the cities. Such distortions are minimal near the equator, but become ever larger as one gets closer to either pole.

An examination of this longitudinal/latitudinal coordinate system provides a way to compare the geometry of the earth’s surface, to compare earth measurement, with the geometry of the flat plane.

The coordinate system that we use for the earth is tied to the four directions, north, south, east, and west. But these four directions are not created equal. The north-south longitudinal lines stretching between the North and South Poles are all straight lines or, more precisely, *great circles*, also known as *geodesics*. They all run in the same direction, i.e., north and south, but they start converging near the poles and ultimately meet at the poles. To run north means to point to the North Pole. On the other hand, the lines that run east and west (latitudinal lines), intersecting the longitudinal lines at right angles (90^o), maintain a constant distance from each other and never intersect each other. But only one of them, the equator, is a great circle, a geodesic. The others, to varying degrees, circle around one of the poles and their curvature becomes pronounced in the vicinity of the poles. To go west or east, on a large scale, is to maintain a constant distance from the poles.

So suppose one starts by going west from somewhere in the northern hemisphere. The instant that one tries to continue in a straight line, i.e., a great circle, one will stop going west. This means that one will leave the latitudinal

circle, one will stop going west. This means that one will leave the latitudinal circle and will begin to veer to the south. (One doesn't recognize this on a flat map for which both longitudinal lines and latitudinal lines are represented as straight lines. But it becomes clearer when one looks at a globe.)

This fact is most clearly and dramatically seen near the North Pole. Suppose, for example, that one begins from a point 10 feet south of the North Pole and, with due determination, heads west. Heading west means maintaining a constant distance from one of the poles. So it means going in a circle around one of the poles: in this case, the North Pole. It means that one keeps the North Pole consistently to one's right. In pursuing this plan, one finds oneself walking in a circle, a circle with a radius 10 feet, around the North Pole.

Now assume, instead, that one walks in a straight line, a great circle, from that same point, initially facing West and 10 feet from the North Pole. In that case, one quickly leaves the North Pole behind as one would leave one's house behind if it happened to sit 10 feet from the street at the start of a journey. In the case of the house, after a couple of blocks the house is almost directly behind the traveler. Similarly, in the case of the North Pole, by the time one has walked 100 yards, with the North Pole now almost directly at one's back, one is, for most practical purposes, heading directly away from the North Pole, that is one is heading south. And if one kept going in a totally straight line, i.e., a great circle, and if the earth were a perfect sphere, one would ultimately miss the South Pole by just 10 feet.

As a final limiting case, if one stands at the North Pole, then no matter which direction one faces, one is facing south.

In sum, the Earth's surface does not admit a grid of straight lines intersecting at right angles. The reason is that, unlike the Euclidean plane, the Earth is not flat and the lines on its surface do not satisfy the Parallel Postulate. On such a surface, one has a choice. One can walk in a great circle. Alternatively, one can maintain a constant distance from a great circle such as the equator. But one cannot, simultaneously, do both.

Although my primary interest has been to characterize the measurement implications of the Parallel Postulate, this example illustrates the force of Playfair's version of the Parallel Postulate (see Chapter 1), as characterizing the flatness of the Euclidean plane.

Perception is the Base

As I pointed out in Chapter 1, in the case of triangles, one's understanding of triangles is needed to understand more complex figures. The same principle applies to the Postulates. For example, the study of geometry on the surface of the earth, of the relationships between places and distances on the earth, requires Euclidean geometry as its base. This is also true for astronomical measurements and remains true even insofar as measurement across astronomic distance, by means of light rays, is nonEuclidean geometry, the geometry of the perceptual level, remains the frame of reference of one's geometric measurements and provides the benchmark to which all relativistic corrections must relate. To make a relativistic correction *is* to relate an observation to the perceptual level.

Both cases, then, exemplify the same principle: that all conceptual knowledge must be related to the perceptually given. Measurement is meaningful *because* it specifies a quantitative relationship to something that one can perceive.

Parallel Lines: The Key Propositions

Euclid's key propositions regarding parallel lines are Proposition 29, and its partial converse, the earlier Proposition 27. The statements and their consequences are far more important than their proofs, which I omit.

Proposition 29 states:

"A straight line falling on parallel straight lines makes the alternate angles equal to one another, the exterior angle equal to the interior and opposite angle, and the interior angles on the same side equal to two right angles."[14]

Proposition 27 states:

"If a straight line falling on two straight lines makes the alternate angles equal to one another, the straight lines will be parallel to each other."[15]

Always remember Euclid's definition of parallel lines: two straight lines in the plane that never meet. The picture below illustrates the content of both propositions:

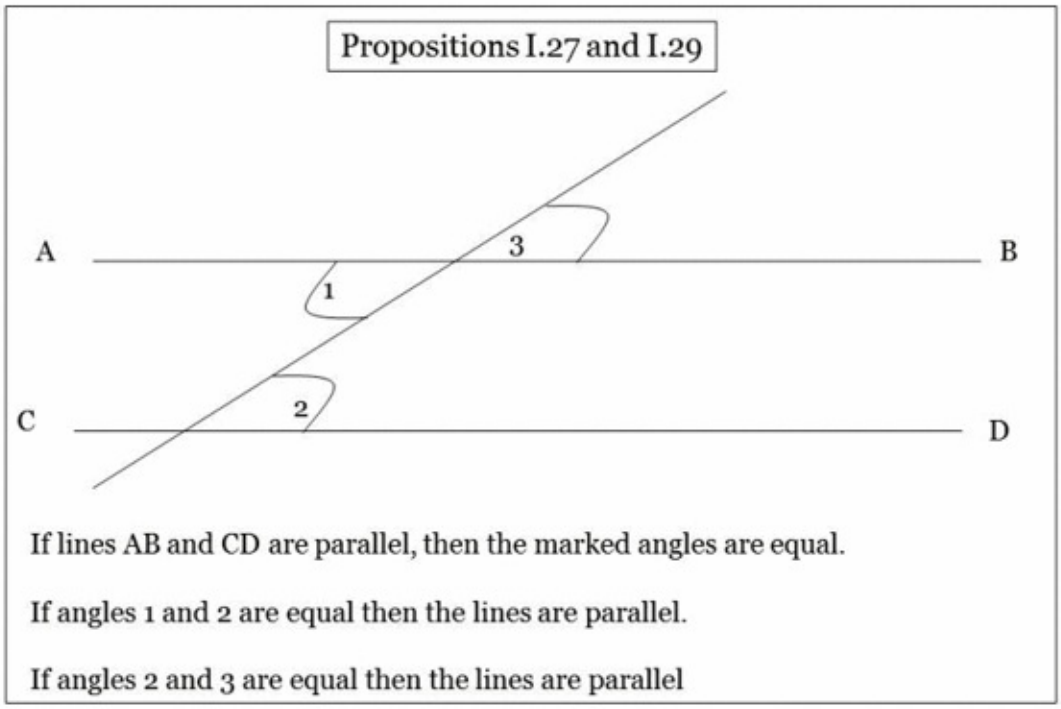


Figure 3

The first statement, that the marked angles are equal when the lines are parallel, was the key insight behind Eratosthenes's estimate of the circumference of the earth.

Beyond its consequences for angles, parallelism has important consequences for lengths, as well. Of particular importance for Euclid's theory of geometric area, are Propositions 33 and 34.

Proposition 33 reads

"The straight lines joining equal and parallel straight lines (at the extremities which are) in the same directions (respectively) are themselves equal and parallel."[16]

The meaning and proof are indicated in Figure 4. The dotted line divides the quadrilateral into two triangles. Relying on the demonstrated properties of parallel lines and of triangles, Euclid shows that the triangles are congruent. This implies, first, that line AB equals line CD and, second, that the corresponding angles at B and C are equal. This second implication establishes that line AB is parallel to CD.

In this argument, once again, one sees the common pattern illustrated in Chapter 1. One performs a measurement, drawing the line from B to D, and then identifies a series of mathematical relationships, to establish the conclusion.

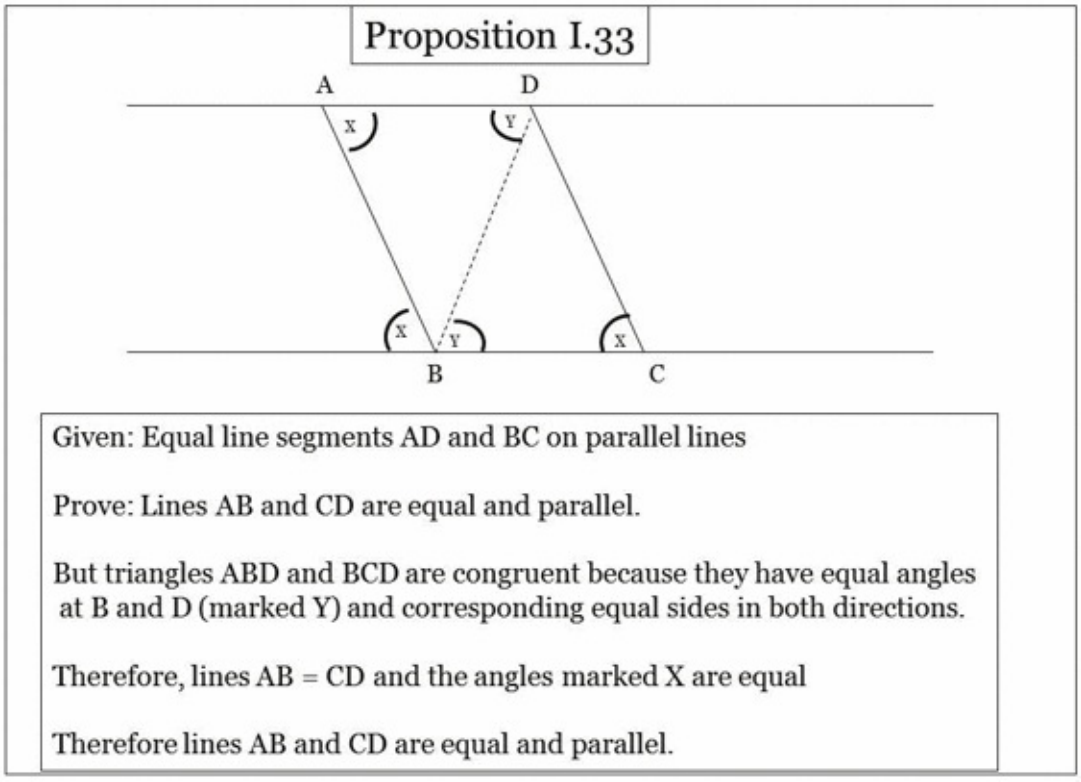


Figure 4

Proposition 34 is a partial converse. It states:

"In parallelogramic areas the opposite sides and angles are equal to one another, and the [diagonal] bisects the areas."[17]

Once again, one draws the diagonal to divide the parallelogram into two triangles. One argues that the triangles are congruent. In light of basic properties of parallel lines, one argues that various angles are equal. One applies Proposition I.26 to conclude that the triangles are congruent. The rest follows from congruence of the triangles:

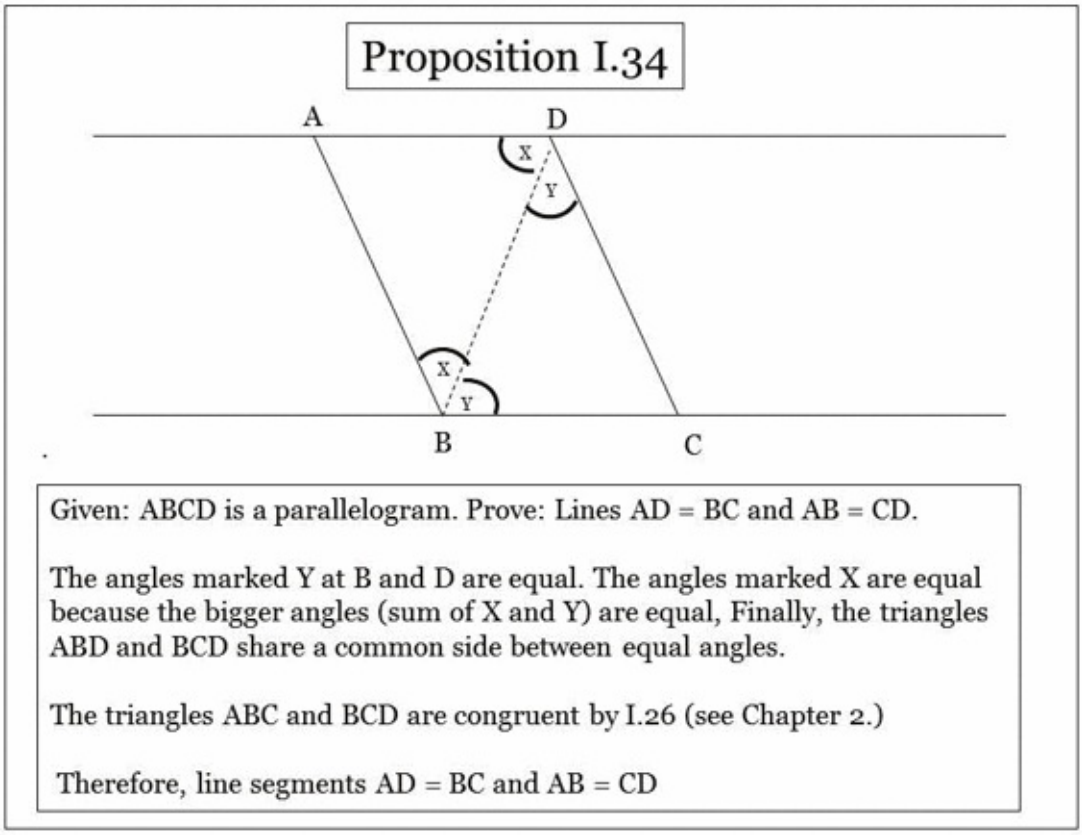


Figure 5

Taken together, these propositions complete one's knowledge of the key properties of parallel lines. And they complete the prerequisites for Euclid's theory of area.

The properties of parallel lines are the foundation of Euclid's theory of areas and volumes. To that development I turn.

Euclid's Analysis of Geometric Area

Chapter 2 took an avowedly geometric approach to numbers. It focused much more on the *quantities* that numbers measure and less on the numerical *expressions* of these measurements. It focused on magnitudes from the perspective of measurement and on measurement from the perspective of its object. But, compared to Euclid, the chapter was very number oriented. For all its focus on the geometric perspective, and, indeed, for all its debt to Euclid, Chapter 2 was a modern perspective, very much *not* in the spirit of Euclid's

Elements, particularly in its analysis of area.

Chapter 2 treated the measurement of area, as it relates to multiplication, and ratio, as a relationship of magnitudes. And Chapter 2 did not apply ratio to geometric proportion. In contrast, Euclid develops his theory of area in exclusively geometric terms and uses it, together with his theory of ratio, to establish his theory of geometric proportion, Euclid's epochal, path-breaking treatment of area and proportion epitomizes, perhaps more than any other aspect of his work, Euclid's distinctive approach to measurement.

The measurement of area as we know it depends, inescapably, on the nature of parallel lines. So, of necessity, my approach in Chapter 2 depends on the nature of parallel lines. But Chapter 2 hides that dependency, even while hiding it in plain sight. The modern approach, which one takes completely for granted, is to count squares or rectangles, as I did in Chapter 2. One knows that rectangles and squares are bound by two pairs of parallel lines. But who would notice that the nature, the very existence, of squares and rectangles reflect the existence and nature of parallel lines?

In this, and other ways, Euclid stands in stark contrast to modern approaches. In Euclid's treatment, parallel lines take center stage.

The Eudoxus/Euclid theory of ratio, offered in Book V, is certainly about magnitudes, but it is not specific to geometric magnitudes (such as length or area) and is independent of the parallel postulate. But the *application* of ratios to similar triangles, Euclid's theory of geometric proportion, flows directly from his theory of area, developed in Books I and II. And the reliance of Euclid's theory of area on the nature of parallel lines is evident from his very first proposition on area. In sum, his entire theory of area derives explicitly from fundamental properties of parallel lines.

The Eudoxus/Euclid theory of proportion is one of the towering achievements of Greek geometry. It is the base of trigonometry, of our ability to measure astronomic and microscopic distances and other geometric relationships. It provides the mathematical foundation for astronomy, navigation, geographic mapping, and all of the physical sciences. We appeal to it whenever we make an architectural drawing, a scale model or scale drawing of any kind. And the entire edifice rests on Eudoxus's theory of ratio and Euclid's theory of area, resting, in turn, on his theory of parallel lines.

Euclid's development of area and of geometric proportion is clear and beautiful in its elegance. But, as I noted in Chapter 1, Euclid had a tendency to focus on the means of measurement at the expense of the object of measurement, focusing, for example on lines and circles without mentioning the directions and distances that they measure. This tendency is particularly noticeable in Euclid's treatments of ratio and area. Thus Euclid expounds Eudoxus's theory of ratio without ever telling us what a *ratio* is. For example, are ratios numbers? Or are they only *sort of* like numbers? Or are they something else entirely?

I provided and explicated Euclid's essential definitions regarding ratio in Chapter 2. I showed just why they make the distinctions that one needs to make, why they make sense, how they relate to the Axiom of Archimedes, and how they relate to a more modern perspective. But, from Euclid's presentation of these definitions, one knows *only* that two ratios can be equal or, if they're not equal, that one is larger than the other one.

Euclid certainly offers a criterion to determine which alternative holds in any given instance. But, one of the ratios is larger? In what respect is it larger? Can one ratio be twice a second ratio? What does it actually mean to be larger? No answer is given. Eudoxus and Euclid knew what problem they were trying to solve and I must presume that they understood how the definitions solved their problem. But Euclid did not articulate their reasons and one is left with apparently arbitrary definitions as the foundation for the most consequential propositions in the entire Euclidean corpus.

Similarly, Euclid develops his theory of area without ever telling us what *area* is. Euclid tells us that two triangles are "equal" in cases where they are clearly not congruent. But he never tells us in what *respect* they are equal. The very word "area" is missing.

In both its virtues and its flaws, Euclid's development of area and geometric proportion epitomize Euclid's distinctive approach to measurement. It is important and instructive to understand that approach, to appreciate both its signal virtues and its limitations.

Euclid's Geometric Treatment of Area

I emphasize in Chapter 1 that Euclid's geometry proceeds without ever selecting a standard of measurement. Yet there is a kind of measurement, an abstract form

a standard of measurement. Yet there is a kind of measurement, an abstract form of measurement, behind every proposition. Every proposition expresses a quantitative relationship and every argument is a chain of abstract measurements that establishes or prescribes a quantitative relationship. In pattern, Euclid argues things like

● two geometric magnitudes (length or angles) are equal, ● one magnitude is greater than another,
● a magnitude is divided into a number of equal pieces, ● or the sum of two magnitudes is greater than another

magnitude.

As an example of this last, the celebrated and important "triangle inequality", Euclid's Proposition 20, states "In any triangle two sides taken together in any manner are greater than the remaining one."[18]

I have made much of the fact that Euclid makes all of these judgments without choosing or even alluding to a standard of measurement except in the case of angles where he really had no choice.

Euclid's approach to area follows the same pattern. Without formally defining or otherwise indicating the quantity he is discussing, Euclid begins his discussion of area without acknowledging or even hinting that he has introduced a new kind of geometric relationship. In effect, Euclid relies on an ostensive definition, as he had implicitly done for length and direction. As Heath, in his notes, puts it, Euclid introduces a new kind of equality.[19] Euclid does not explain what this new equality is comparing and, in this respect, his treatment is a puzzle left entirely to the reader. Nonetheless, once one grasps that he is talking about area, one can follow his reasoning. Euclid's discussion covers enough ground that, by the end of Book I, without ever producing a formula for area, Euclid was able to prove the fundamental Pythagorean Theorem as expressing a relationship between the areas of certain squares.

Book II developed the theory of area further, but there is a limit to what Euclid could say without a theory of ratio. Later, having developed a theory of ratio in Book V, Euclid returned to complete his account of area, among other related topics, in Book VI and provided the underpinnings of the modern formulas for area.

Euclid's treatment of area is no longer in the curriculum. The modern focus is on formulas involving numbers that one attaches to lengths, areas, and volumes. One is asked to memorize formulas for the areas of rectangles, triangles, parallelograms, circles, and the surface of a sphere, with or without understanding these formulas. If an attempt is made to justify, say, the formula for the area of a rectangle, the standard approach is to count squares, as I did in Chapter 2 in my discussion of magnitude. If this were done before the formula is provided to be memorized, and if some attempt were made to lead the student to such an approach, counting squares is a good way to look at measuring area. But something is still lost if one doesn't *also*, at some point, look at it from Euclid's perspective. Measurement is a form of *identification* but in one's haste to apply a number it is too easy to lose sight of the reality that these numbers are used to measure.

(And the worst thing one can do is to simply start with a formula, treating it as though nothing beyond simply memorizing it were necessary, aborting the process of understanding before it can begin.)

Its shortcomings, notwithstanding, once one grasps that Euclid is discussing *area*, one never loses sight of the geometric property under investigation.

Euclid begins with Proposition 35, which reads:
"Parallelograms which are on the same base and in the same parallels are equal to each other."[20]

Equal in what respect? Euclid is talking about area, even though he never tells us. Most importantly, he is *not* talking about equality of numbers.

The argument is outlined in Figure 6. Euclid, ingeniously, creates a figure that contains both parallelograms. He then subtracts, in turn, two distinct triangles from the composite figure leaving, alternately, the respective parallelograms. Since the triangles that he subtracts are *congruent* and, therefore, have the same area, the remainders, the two parallelograms, have the same area. Only Euclid says nothing about area; he doesn't give us the word and the reader is left on his own to infer the concept.

The triangles are congruent because of the properties of parallel lines. Parallel lines are essential to Euclid's development of areas and the appeal to these properties of parallel lines begins with his first proposition (i.e., Proposition 35) concerning areas.

In what follows, I will quote Euclid's propositions as they stand in translation. But I will, in my *discussions* of Euclid's arguments, supply the word, *area*, that Euclid leaves out.

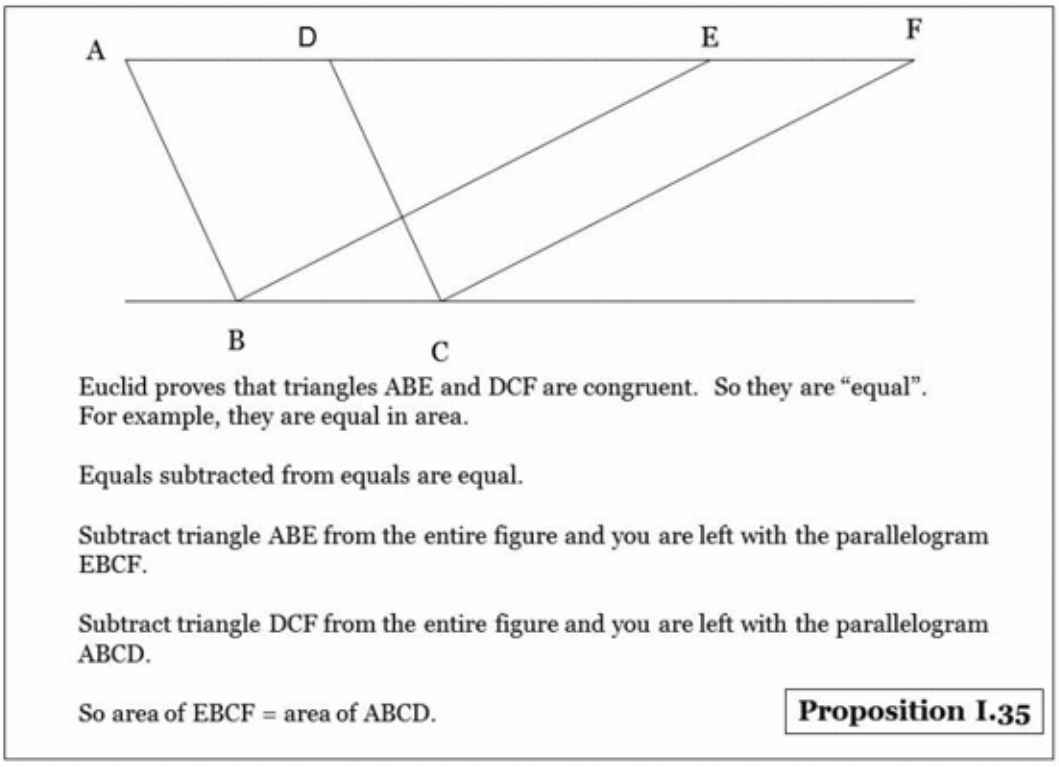


Figure 6

Interestingly, Euclid has not demonstrated a way to cut one of the parallelograms into pieces to reassemble one into the other. Yet his demonstration, if one follows it, is perhaps more convincing than if he had. His recipe works by cutting something out of a bigger figure. If one cuts out a triangle in one way, one gets the first parallelogram; if one cuts out a different triangle, *of the same shape and size*, another way, one gets the second. The remainders must be equal in both cases.

"Equals subtracted from equals are equal," is the pivot point of an argument that could easily be taken as either meaningless or equivocal. It is a statement that pertains to magnitudes, not to shape. Yet up until now, Euclid's comparisons of shape have either involved congruence or have focused on various parts of the shapes: the edges and angles.

In this case, the figures that Euclid, alternatively, subtracts from the larger shape have exactly the same shape and, therefore, are equal in all geometric respects.

including area. The figures that alternatively remain have different shapes. But, one realizes, their areas are equal. When, and if, it occurs to one that Euclid is talking about area, one realizes that the apparent equivocation, properly understood, is not an equivocation.

Euclid is making measurements in the general sense that I have called abstract measurement and he is doing so without anything more than an ostensive definition of what he is talking about. Indeed, as ostensive definitions go, this one is completely implicit. Euclid is measuring … something. And there is nothing else that he could be measuring, that could be equal in the two alternative shapes, but the area of those shapes.

I pointed out in Chapter 1 that Euclid focuses more on lines, angles, and circles than on what he uses them to measure. Four of his five postulates measure direction, yet his closest use of the concept, *direction*, is to identify the two directions along a line from a point on that line.

Euclid follows the same pattern with area. He does not introduce the word "area," he merely maintains that two triangles are equal or that two parallelograms are equal. He does not name the respect in which they are equal. He simply introduces without notice, as Heath puts it, "a new conception of equality between figures."[21]

Proposition 36 takes the next step:
"Parallelograms which are on equal bases and in the same parallels are equal to each other."[22]

Again, the picture (Figure 7) tells the story. Euclid leverages Proposition 35, the properties of parallel lines, and the common notion, "Things which are equal to the same thing are equal to each other."

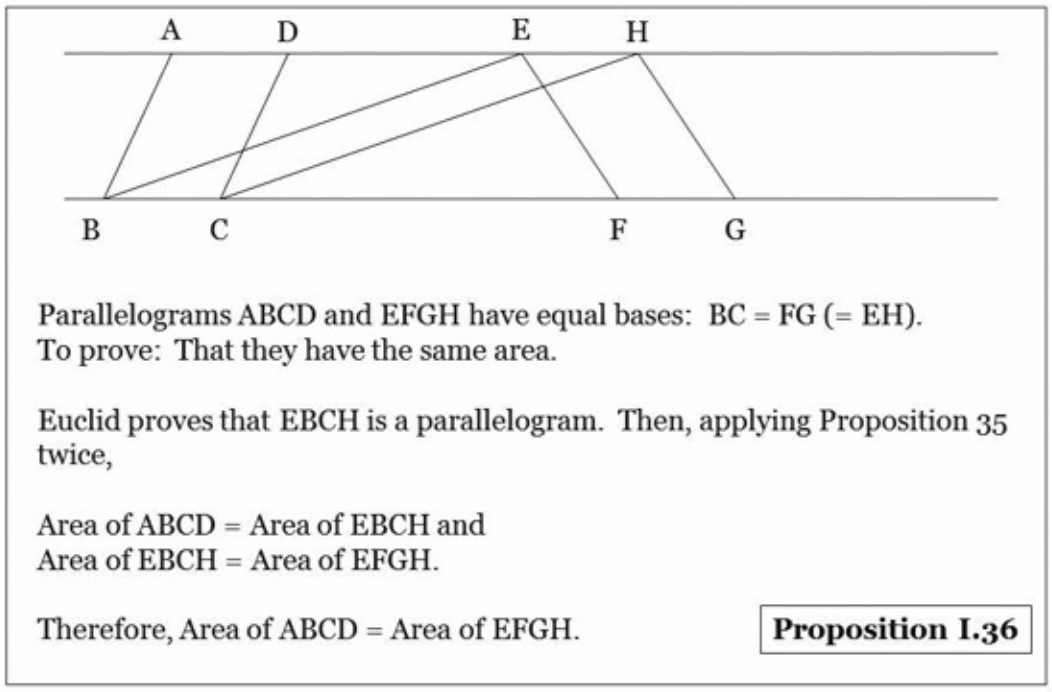


Figure 7

Once (and if) one grants and grasps Proposition 35, the rest is easier. Euclid continues in the same vein, proving, for example, Proposition 38:

"Triangles which are on equal bases and in the same parallels are equal to each other."[23]
And Proposition 41:

"If a parallelogram have the same base with a triangle and be in the same parallels, the parallelogram is double of the triangle."[24]

So far, Euclid can compare areas of parallelograms or triangles only when they fit between the *same* pair of parallel lines and have their base on one of those lines. It's a start, but only a start. However there is a surprisingly simple device that breaks down this barrier and it's based upon Proposition 43: "In any parallelogram the complements of the parallelograms about the diameter are equal to one another."[25]

Figure 8 shows what this means:

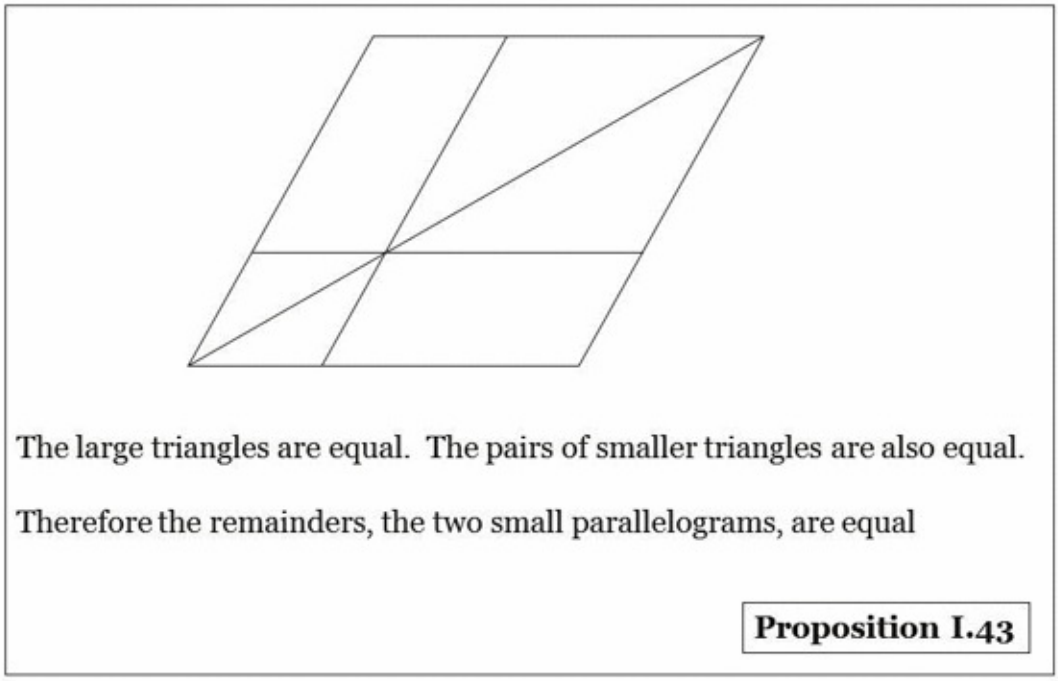


Figure 8

Euclid is still subtracting equals from equals, but, by now, one knows what he is measuring.

Proposition 43 is important because the two equal parallelograms are neither the same shape nor do they lie in the same parallels. And it provides the key to the following problem: Given a parallelogram and two parallel lines, find another parallelogram, one of equal area, between the two parallel lines. Figure 9 details the construction:

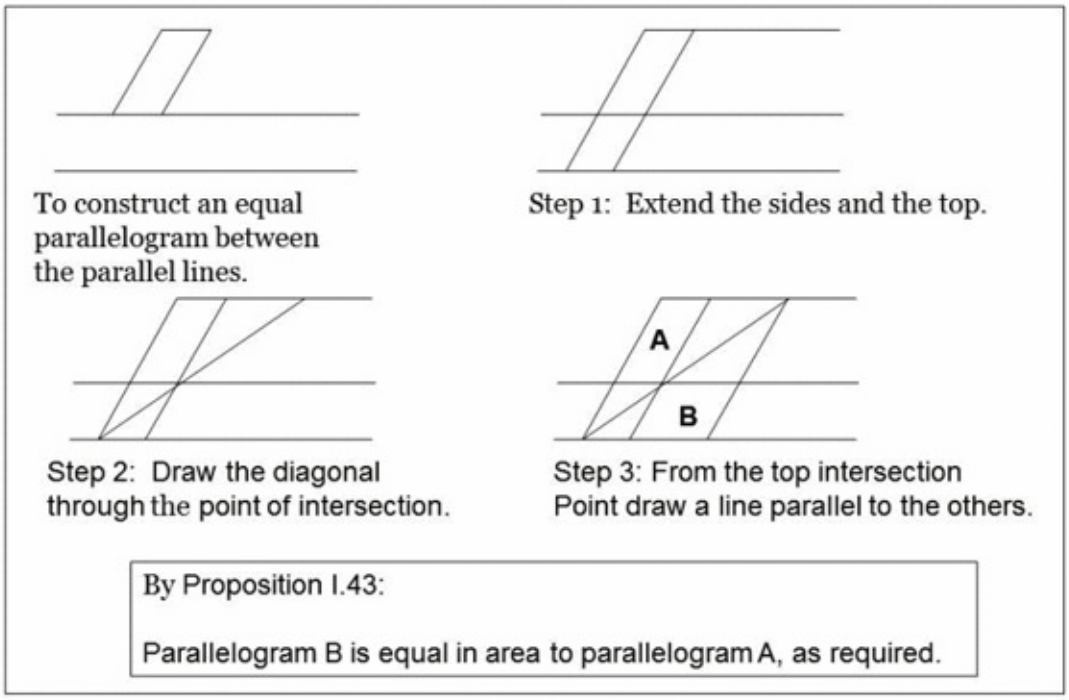


Figure 9

With this result, Euclid still doesn't have a formula, but he has found a geometric process, an abstract measurement, for comparing, by construction, the areas of any two rectilinear figures.

Book I, on the basis of his propositions on measuring area, culminates in the fundamental Pythagorean Theorem, Proposition 47:

"In right-angled triangles the square on the side subtending the right angle is equal to the squares on the sides containing the right angle."[26]

As I have noted earlier, the Pythagorean Theorem is not, for Euclid, a formula relating the *sides* of the triangle; he does not have such a formula. The proposition offers, rather, a formula about equality of areas, that the square on one side, the hypotenuse, is equal to the sum of the squares on the other two sides. To emphasize further: This is *not* a sum of *numbers*, but, specifically, of *areas*.

Euclid's conclusion derives from a rather long series of equalities, but each such equality is established by way of direct comparisons of geometric shapes with the aid of Euclid's Common Notions.

At this juncture, a picture will be helpful to show specifically what the Pythagorean Theorem is saying.

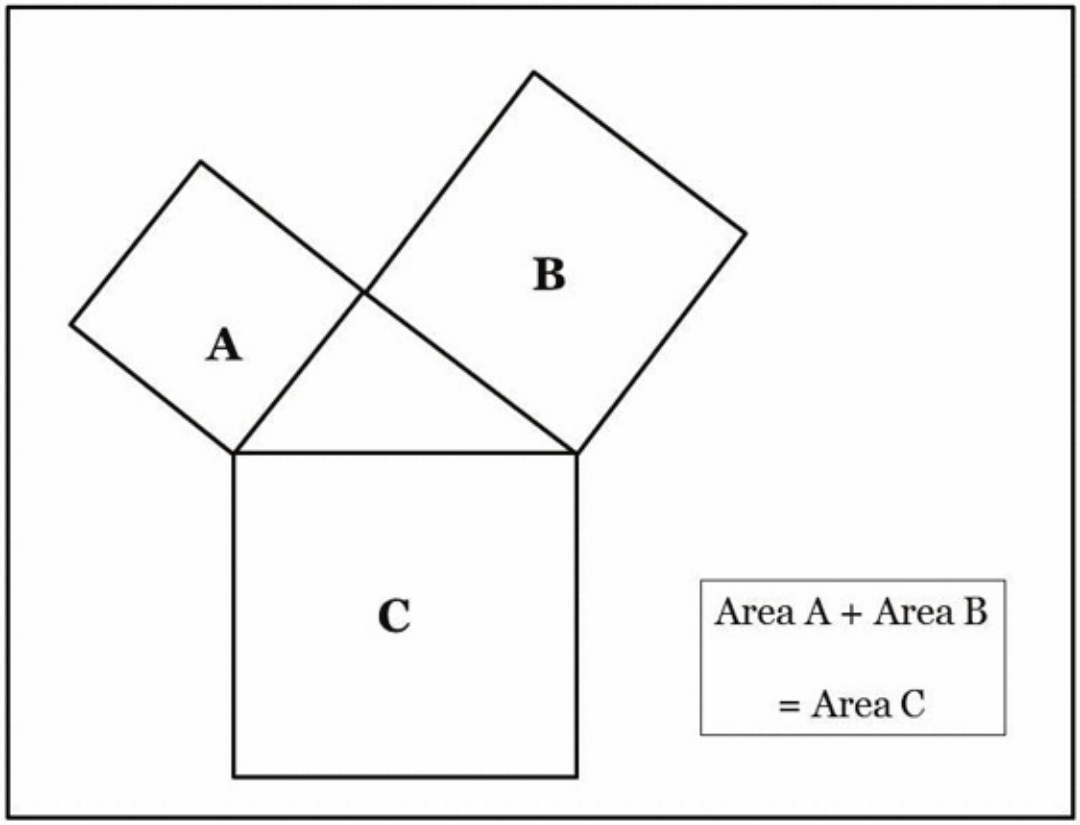


Figure 10

The triangle in Figure 10 is a right triangle with a right angle at the top. The letters *A, B, and C represent areas of squares*. Euclid has developed ways to compare areas, but has not provided a way to quantify them, to attach numbers to them. Nonetheless, his propositions to this point are sufficient to argue that

Area A + Area B = Area C

Euclid's proof is difficult to follow. But in the Euclidean spirit, I offer a well-known alternative that makes the theorem almost obvious visually. It's roughly in Euclid's spirit because it only depends upon re-arranging various geometric shapes. In Figure 11, compare the first figure on the left with the last one on the right. If the four triangles are removed from each, the remainders are the square on C, in the first figure and the squares on the other two sides (A and B) in the last figure.

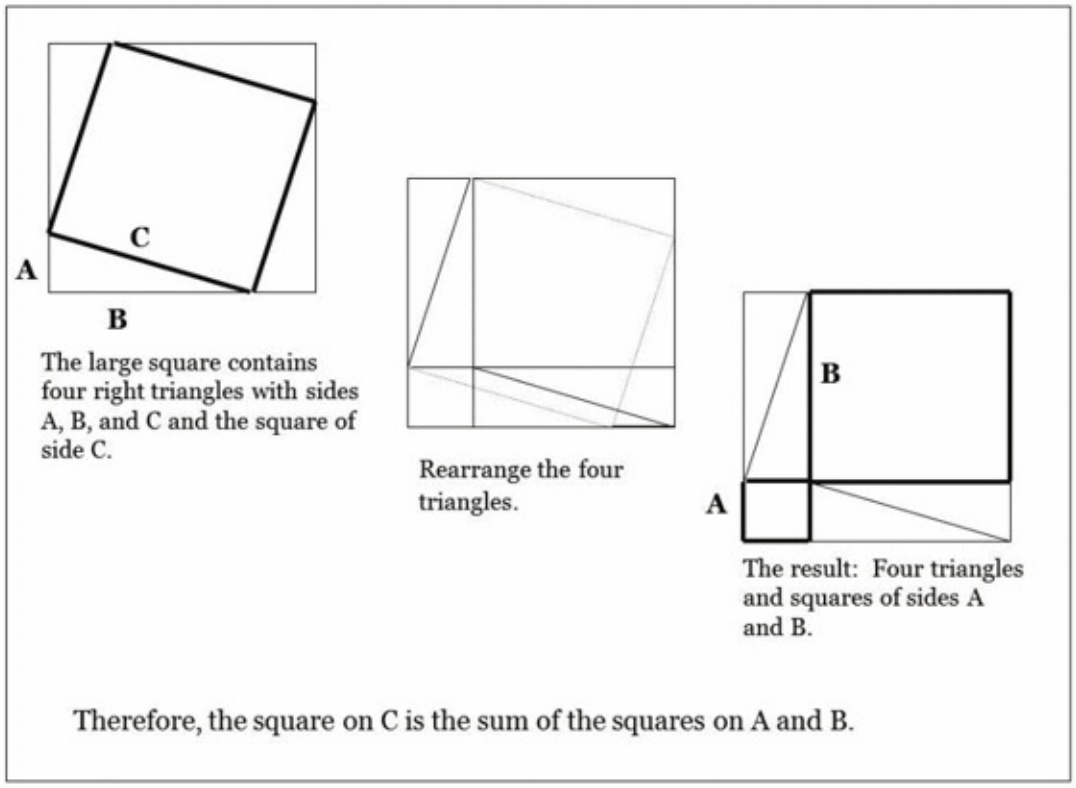


Figure 11

Like Euclid's proof, but more transparently, this figure expresses the meaning and demonstrates the truth of the Pythagorean Theorem by means of a process of abstract measurement. In general, Euclid compares figures by direct and indirect means according to their equality or inequality, establishing his quantitative relationships by a series of such comparisons. In its deepest sense, this is the essence of measurement, considered as a process.

As a final comment, it is well known that the validity of the Pythagorean Theorem depends upon the Parallel Postulate and the properties of parallel lines. Thus, it is appropriate that it comes out of Euclid's treatment of area which, itself, depends upon the Parallel Postulate.

Book II continues the study of area, carrying it as far as Euclid can without a theory of ratio. From a general point of view, the most interesting Proposition provides the appropriate extensions of the Pythagorean Theorem to the more complex statement that applies to more general triangles.[27] (See also Chapter 7.) But, for my present purposes, the most interesting Proposition is the last, Proposition II.14, which I quote without proof:

"To construct a square equal to a given rectilineal figure."[28]

Prior to this final conclusion, Euclid could find a rectangle, of any prescribed height, equal in area to a given rectilineal figure. With II.14 he can construct the unique equal rectangle that is also a square. Yet he still doesn't have a formula or a way of identifying an area numerically!

For Euclid, finding a square equal to a given area *is* measuring the area. Indeed, when the Greeks spoke of "squaring the circle," they were asking for a way to construct a square that would have the same area as a circle. The problem, as posed, as being executed by straight edge and compass, cannot be solved. It was left for Archimedes to do the next best thing: He proved that the area of a circle was equal to that of a right triangle with one leg equal to the circumference of the circle and the other equal to its radius.[29]

In this, Archimedes looked both forwards and backwards. He looked backwards in the way that he specified his answer. He looked forwards in using a limiting process (Eudoxus's method of exhaustion) to measure something that straight edge and compass could not. Archimedes could *specify* the dimensions of the triangle; but he could not construct it.

Euclid's strengths were his focus on the geometry and his systematic, if only implicit, use of abstract measurement to reach his conclusions. His essential weaknesses, beyond the Platonic elements of his work, were two. The first, introducing fundamental quantities without even naming them, I have already mentioned. The other was his steadfast avoidance of units. In order to provide a formula, e.g., a formula for area, one must multiply numbers. But even before that, one must have assigned numbers to lengths and, continuing the example, to assign a number to a length, one must first choose units.

Even in Book VII, in which Euclid presents number theory (illustrated to look like geometry) Euclid only goes half way. In Book VII Euclid assumes a unit, a specific magnitude that divides every other magnitude under discussion. By intention and explicit definition, Euclid treats his unit as the number one or, more precisely, as something that is "called one."[30] Still, the principal use of units in Book VII, for example in his proof of VII.2, the famous *Euclidean Algorithm*, is to insure that various repeating processes must terminate.[31] It is not, as one might have expected, so that he can assign numbers to lengths.

More generally, Euclid does not motivate his concepts. As one early important example, at the beginning of Book I, he offers a definition of parallel lines without providing a reason to think that pairs of lines satisfying his definition actually exist and without explaining why one should care or what the concept measures.[32] In the case of area, Euclid omits to even name the concept and leaves the reader to figure out, on his own, what Euclid is talking about.

Euclid, unlike the moderns, does not measure area by counting squares. He does not provide a way to use numbers to identify area. Yet measuring relationships involving areas was important to him; first, as another aspect of measuring area, but also because his theory of geometric proportion depended on it. To this subject I turn.

Ratios and the Theory of Geometric Proportion Introduction

Book V presents Euclid's theory of ratio (after Eudoxus) and lays the foundation for Book VI. I will not repeat the elucidation of Euclid's key Book V definitions that I provided in Chapter 2. However, it is worth pausing to understand what was at stake; why the complexity of Book V was necessary, why solving the problem of ratio was so important, and just in what ways Euclid's *presentation* of the solution (a solution due, at least in part, to Eudoxus) was so typical of Euclid's approach.

The power of indirect measurement is a major legacy of Euclidean geometry and its crown jewel is the theory of geometric proportion, presented in Euclid's Book VI.

The theory of geometric proportion is at least as old as Pythagoras. But the Pythagorean theory of geometric proportion had collapsed, ironically by virtue of two path-breaking discoveries by the Pythagoreans themselves. One of these was the celebrated Pythagorean Theorem relating the sides of a right triangle to its hypotenuse. And the other was the finding that$\sqrt{2}$ is irrational, that the diagonal of a square is incommensurate with its sides.

Beyond the bare outlines of this story, little is known definitely; the Pythagoreans left no direct record of their work. The knowledge of the Pythagorean Theorem and the irrationality of$\sqrt{2}$, were not lost, and a satisfactory

demonstration of geometric proportion was left as a challenge for future geometers. But one can only guess at the Pythagorean approach to geometric proportion and any guess is speculative.

Nonetheless, it is easy enough to speculate and I offer one possibility, not as a serious hypothesis, but only as a way of illustrating the kind of approach one might have taken prior to the work of Eudoxus. My purpose in doing so is, by shedding some light on the kind of issue that the Pythagoreans faced, to better appreciate the contributions of Eudoxus and Euclid.

The Commensurate Case

Book VII of Euclid's Elements is thought to contain the earlier Greek approach to measuring proportion, the approach to ratio that predates the discoveries of Eudoxus presented in Euclid's Book V. So I start with the essential definitions for Book VII. Definition 20 in Book VII reads:

"Numbers are **proportional** when the first is the same multiple, or the same part, or the same parts, of the second that the third is of the fourth."

By way of further explication:

Definition 1: "An **unit** is that by virtue of which each of the things that exist is called one." Definition 2: "A **number** is a multitude composed of units." Definition 3: "A number is **a part** of a number, the less of the greater, when it measures the greater;" Definition 4: "but **parts** when it does not measure it."

In modern terms, if A, B, C, D are the four numbers involved, Euclid's concept of proportional amounts to A/B = C/D. But, though Euclid speaks of numbers, what he shows is line segments. These segments serve to represent numbers, according to Definition 2, because they are all considered to be multiples of some common unit a length "that is called one."

And notice that it is not obvious what Euclid is actually referring to. For example, is *5* a number? Or is a *collection* of 5 *things* a number? If Euclid draws a line segment that is 5 times the length of another line segment designated as the unit, is the line segment a number because Euclid thinks of it as a multitude of 5 units? Or is it just a *line segment* that happens to be, and is taken to be, 5 times the length of his chosen unit. Clearly, today, we would say that 5 is a

number and a multitude is something like a collection for which numbers are used to count. But Euclid's formulation says that the number is the *multitude* (the 5 *things*), or, in Book VII, the *measured length*, not the *count* (5) of the multitude. According to Euclid, as I take it, *5* is not a number. Rather, a *collection* of 5 marbles *is* a number, each individual marble being called one.

In Book V, Euclid speaks of equal ratio. In Book VII, he does not. Rather, he speaks of two pairs of numbers being *proportional*. His generic term, "parts", is used descriptively in relation to a common measure; and I take the use of the two different terms, proportional and ratio to be deliberate.

As I read Euclid, if A = 3 and B = 7, he might say that A is 3 parts out of 7. Yet if C = 6 and D = 14, he would say that A, B, C, and D are proportional, i.e., that 3 is the same parts out of 7 as 6 is of 14. For this to make sense, one needs to imagine that each part of 14 consists of two units, so that 14 has seven 2-unit parts and 6 includes three 2-unit parts. So, if a single part of 14 contains 2 units, then 6 is 3 parts out of the seven 2-unit parts of 14. Again, in modern terms, Euclid is essentially saying that if each pair is reduced to lowest terms, then the corresponding terms resulting reduced fractions are identical.

I offer these translations purely to relate the Greek concept to our modern perspective. However, the Greeks did not look at it the way we do. Rather, they thought, implicitly, of a *relationship* between two numbers. And that relationship was specified when one could find the largest common divisor, the largest number dividing both numbers, or, in Euclid's terms, "greatest common measure". So, if A = 6 and B = 14, then the greatest common measure of A and B would be 2. So, taking each repetition of the common measure as a "part", one would say that B has 7 parts and that A is 3 parts of 7.

Euclid's Book VII, Proposition 2 offers an algorithm (a series of repeatable steps known today as the "Euclidean Algorithm to find this common measure and this algorithm is the foundation of Book VII.

It is the existence of a common measure that makes two segments commensurate. Simply put, two line segments are commensurate when the Euclidean algorithm terminates. At that point one has found the greatest common measure. If the process never ends, as in the case when A is the side of a square and B is the diagonal, then there is no common measure and the two line segments are incommensurate. This is not an issue in Book VII because

Euclid there assumes throughout that all of his segments represent numbers, that is, that they are all multiples of a particular line segment called the “unit” and representing 1. Numbers, as multiples of one, are always commensurate and the Euclidean Algorithm will always terminate. The unit will not, in general, be the *greatest* common measure, but it is always *a* measure.

Now, how might all of this apply to similar triangles? Why might the ability to find a common measure be relevant to geometric proportion?

Suppose two triangles of the same shape are given and that one can find a common measure of the base of each. In this case, one demonstrates that their sides are in geometric proportion in three steps. Step 1:

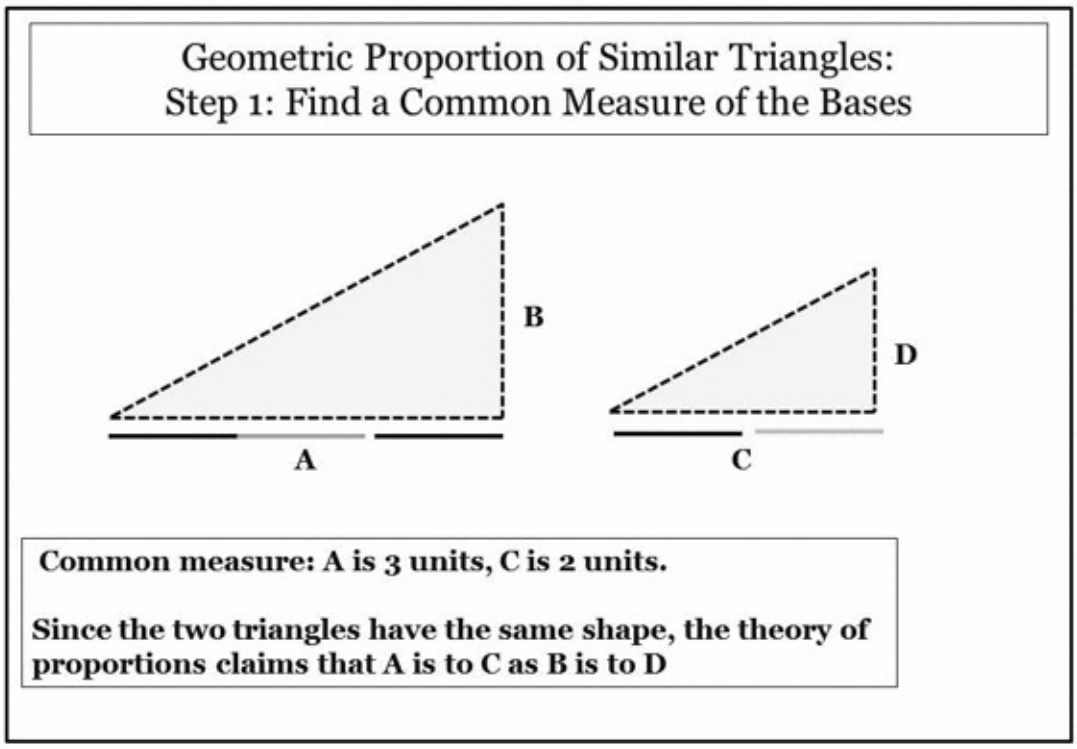


Figure 12

Step 2 is to use the common measure of the respective bases to cut off congruent triangles from each:

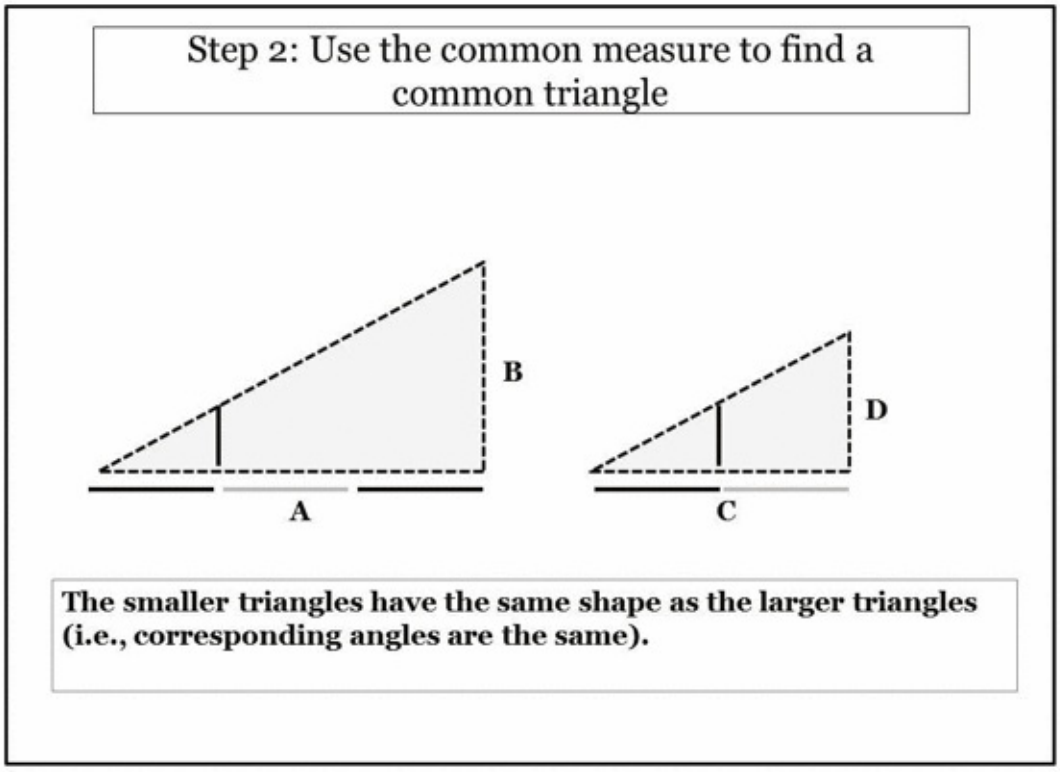


Figure 13

Step 3 is to fill each triangle with the smaller triangles:

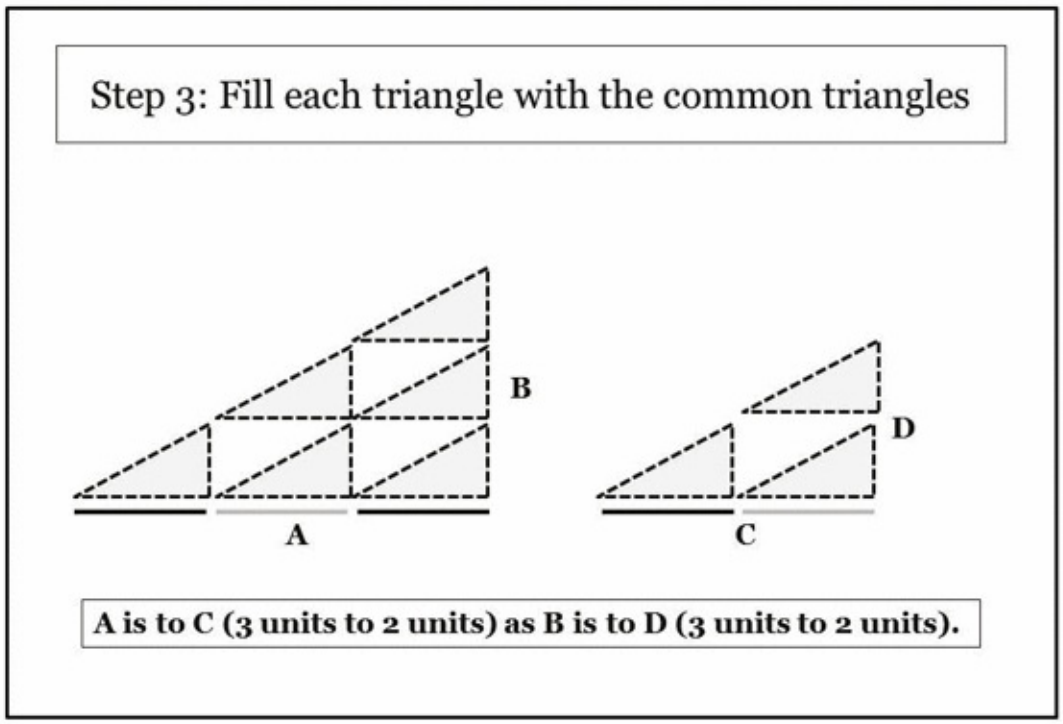


Figure 14

In Step 3, as one adds each column of smaller triangles, successively to the right, within each triangular figure, notice that every additional column has one more

level of small triangles than the previous column. So, in this construction, the number of rows within each figure will always be equal to the number of columns. It follows, that A and B will have the same number of divisions and that C and D will have the same number of divisions. Also, the size of the divisions of A are the same size of the divisions of C and, similarly, for B and D. It follows that A is to C (3 horizontal units to horizontal 2 units) as B is to D (3 vertical units to 2 vertical units). Which is to say that A, C, B, D are proportional. That, in the language of Euclid Book V, the ratio of A to C is equal to the ratio of B to D. This is geometric proportion.

The essential point is that this analysis required finding a common measure between A and C to get off the ground. When this is possible between two lengths, these lengths are said to be commensurate. If not, the lengths are incommensurate.

In modern terms, once again, if two lengths are commensurate, their ratio is a rational number. If they are incommensurate, their ratio is an irrational number.

Pythagoras had assumed, and needed to assume that one could always find a common measure. Fundamentally, before the work of Eudoxus, to even *speak* of equal ratio, one had to count units. To speak of A being parts of B, you had to find a common measure of both segments, at which point you could count the instance of that common measure in each segment. Until one could speak meaningfully of equal ratio in incommensurate line segments, it was not even possible to *formulate*, much less *demonstrate*, a general statement of equal proportion for similar triangles.

When Pythagoras famously said, “All is number,” he stated his major premise. He was essentially saying something to the effect that all relationships can be reduced to number, are, at bottom, numerical, reducible, indeed, to whole numbers (positive integers). And this would certainly include the view that every pair of lengths has a common measure. The subsequent discovery of a counter-example, the discovery that the square root of two is irrational, that the diagonal of a square is incommensurate with its sides, struck a blow to the heart of the Pythagorean approach.[33]

The Eudoxus/Euclid Theory

Now consider the first Proposition in Book VI:

“Triangles and parallelograms which are under the same height are to one another as their bases.”[34]

Figure 15 illustrates Proposition VI.1 for triangles sharing the same apex. As we saw (in Proposition I.38), when the bases are equal, the areas are equal. As a result, if the base of one triangle is three times the other, then it has three times the area. The same reasoning applies to subdivisions of the base and, by a further implication, to any two triangles for which the respective bases are related by a ratio of whole numbers. Finally, although my picture shows triangles that share the same apex, the only thing that matters to the conclusion is that they have the same height, that they are bounded by the same pair of parallel lines.

But what if the bases of two triangles are incommensurate? The existence of this possibility is the reason for Euclid’s Book V; the reason that the all-important theorems on proportion can only appear after the developments in Book V.

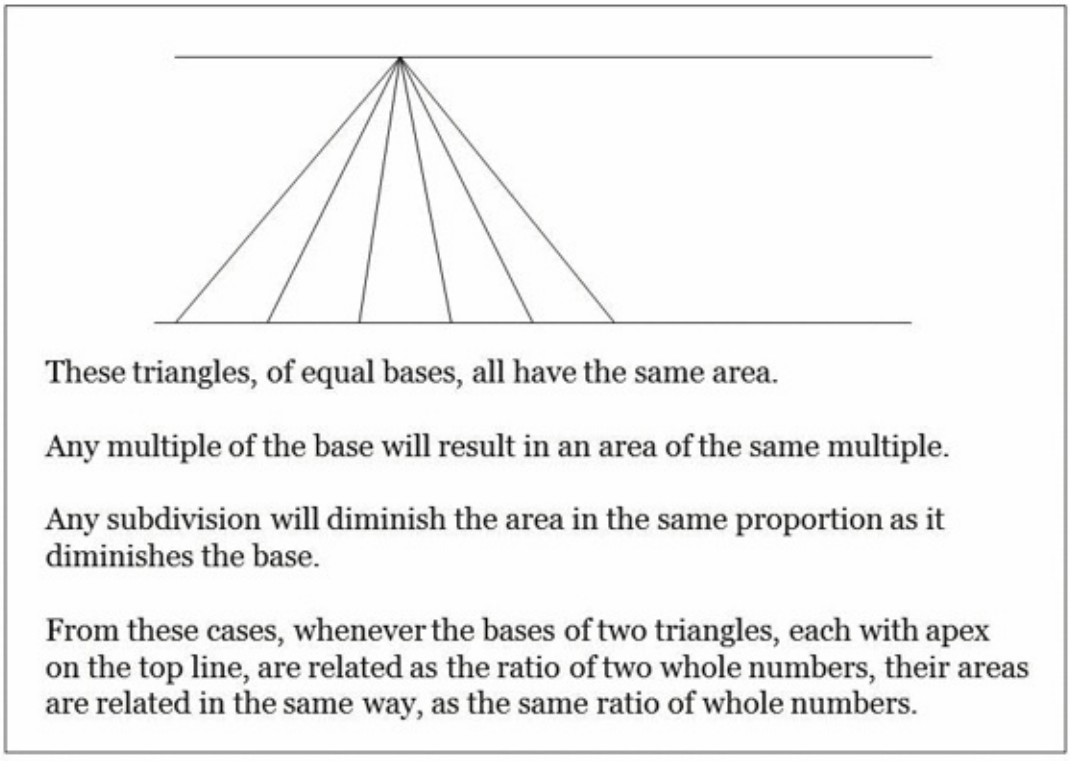


Figure 15

Euclid used Eudoxus’s definition of equal ratio to present a new theory of proportion, not subject to this limitation. Euclid believed that the meaning of *equal* ratio coincided with the previous concept of proportional for commensurate magnitudes. We know this because, he tells us so in Book X,

proposition 5. (His proof of that proposition is rightly considered incomplete, but not fatally so.)[35] Presumably, Euclid was aware, as well, that his definition of *greater and lesser* ratio also coincided with the older definition when applied to commensurate magnitudes, that greater ratio would correspond to more parts, and lesser ratio to fewer parts. In the case of *ratio*, unlike the case of *area*, Euclid at least gives us the word (*ratio*) In Book V. But, as with his conception of area, Euclid does not tell us what a ratio *is*.

Euclid's definition (from Eudoxus) of equal ratio made sense to him, made sense to Archimedes, and, perhaps, made sense to some others. But its formulation is notoriously obscure. Only in the late nineteenth century was it re-introduced, albeit in a different form, as Dedekind cuts.[36]

For all the apparent arbitrariness of these definitions, Euclid accomplished something very important: He told us how to *compare* ratios. He told us when two ratios are equal and when one ratio is greater than another. Ratios express the relationship of two magnitudes, so his definition determines when two *pairs* of magnitudes are related in the same way, have the same ratio, and when one pair of magnitudes has a greater ratio than the other.

This is not a completely satisfying solution, but it solves the problem it was intended to solve, making possible the all-important theory of proportion in Book VI. And its limitations are of a piece with the rest of the *Elements*. Euclid treats ratios the same way he treats distances, angles, and, certainly, areas. In all these cases, he starts out with an ability to make comparisons of equality and comparisons of greater and lesser. For distances and angles, this ability is conferred by the Postulates. For areas, he sneaks it into a Proposition; the definition of area, at best, is implicit and is, essentially, ostensive. For ratios of magnitudes, Euclid offers the apparently arbitrary Definitions V.5 and V.7 in Book V, definitions that do not tell us what a ratio is or how it relates to a ratio of two numbers; but only when two ratios are equal or when one of them is larger than the other. And yet, these definitions happen to be exactly what Euclid needed, are essentially correct, and they formulate the required conditions as well as one could possibly formulate them within the confines of the conceptual framework available at that time.

In all such cases, one can be frustrated with Euclid's approach. But, as I said in the case of area, in taking this approach, Euclid is closer to the actual subject of his enquiry than modern treatments that tend, much more than Euclid, to present

prepackaged definitions as though they had sprung fully formed from the head of Zeus, leaving their lineage, their precise tie to the world, shrouded in mist.

Euclid has much to teach us beyond the actual content of his masterpiece and, perhaps, his basic concepts should be even more ostensive, more focused on their referents in the world, rather than less. Mathematicians still remember how to arrange deductive systems. But they have mostly forgotten how (or, at least, why) to link those deductive systems to those aspects of the world that they were once designed to capture.

To recall, Euclid's Definition 5 of Book V, regarding equality of ratio, reads: "Magnitudes are said to be in the same ratio, the first to the second and the third to the fourth, when, if any equimultiples whatever be taken of the first and third, and any equimultiples whatever of the second and fourth, the former equimultiples alike exceed, are alike equal to, or alike fall short of, the latter equimultiples respectively, taken in corresponding order."[37]

In Chapter 2, I explicated this definition in detail. To summarize that discussion, let A represent the ratio of the first magnitude to the second and B represent the ratio of the third to the fourth. From the modern perspective (though not from Euclid's) 'A' and 'B' may be irrational *numbers*. Then Euclid's definition translates to the modern perspective, as follows:

A = B if and only if the following is true: If n and m are *any* whole numbers, then A is greater than n/m only if B is; it is equal to n/m only if B is; and it is less than n/m only if B is.

Euclid's criterion manages to say this without ever having to recognize a ratio as a number. His criterion requires only the ability to add magnitudes together, to take multiples of them.

The force of this in application is the following: generally speaking, if the equality of two related ratios can be demonstrated to hold whenever the related magnitudes are commensurate, Euclid's definition implies that it holds, as well, when the magnitudes are incommensurate. This circumstance, exploited in Proposition VI.1, is also later exploited by Archimedes, for example, in proving his celebrated law of levers.

The argument for Book VI, Proposition 1 will illustrate the general pattern. I state Proposition VI.1 again:

“Triangles and parallelograms which are under the same height are to one another as their bases.”[38]
Figures 16 - 18 outline the general argument for triangles, appealing to the Definition V.5 definition of equal ratio:

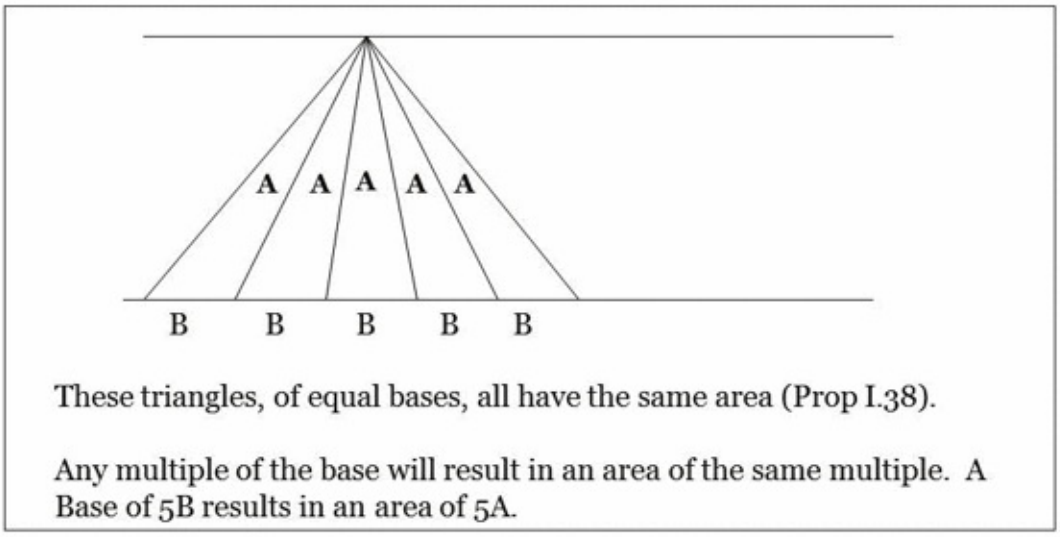


Figure 16

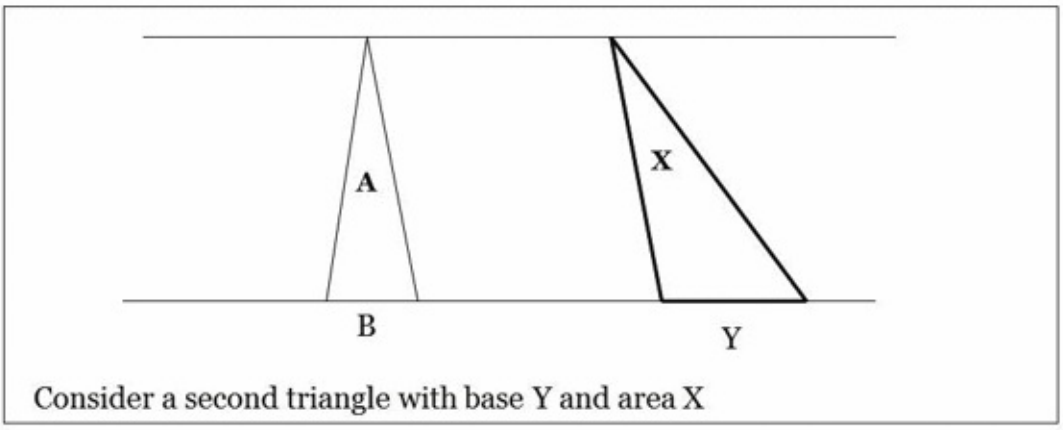
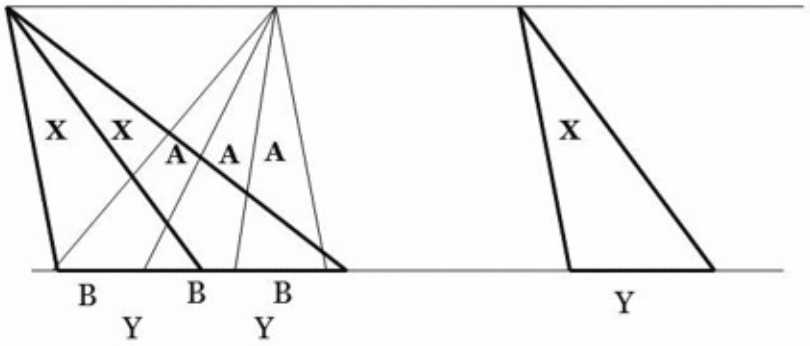


Figure 17

Now apply Euclid’s definition of equal ratio (Definition V.5):



Take a multiple (say 3) and apply that multiple to the first triangle having base B and area A.

Take a second multiple (say 2) and apply that multiple to the second triangle having base Y and area X.

If 2Y > 3B then 2X > 3A because when the base is larger, so is the area.

No matter what multiples one chooses, the two relationships will always be the same. Either both will be '<', both '=', or both '>'. Therefore, by Euclid's Definition V.5, the ratio of B to Y equals the ratio of A to X.

Figure 18

Euclid's criterion applies *because*, as I have already shown, the area and base are proportional whenever the bases are commensurate. The area on a base of 3×B is 3×A; the area on the base of 2×Y is 2×X. If one accepts Euclid's definition and if one sets up the argument properly to prove the general case, it is enough to offer a direct proof for commensurate magnitudes. If one understands this example, one understands the force of Euclid's definition.

It is important to notice that Euclid's definition of equal (or unequal) ratio requires that the first and second magnitudes be of the same kind and that the third and fourth magnitudes be of the same kind. A ratio, for Euclid is always between magnitudes or quantities of the same kind. But, of critical importance, his definition *does not require* that the first and second magnitudes be the same kind of magnitude as the third and fourth. And this is the key point, because Euclid's argument for Proposition VI.1 depended entirely on the ability to compare respective ratios among different kinds of magnitudes. One ratio is a ratio of lengths, the bases of the respective triangles; the other is a ratio of areas, the areas of the respective triangles.

It is astonishing that Euclid's entire theory of geometric proportion, developed in

Book VI *hinges* upon the fact that the first and second magnitudes *need not be the same kind of magnitude* as the third and fourth for their respective *ratios* to be comparable. It is almost as surprising that Euclid compares, and, seemingly, *needs* to compare, ratios of *lengths* to ratios of *areas before* he can compare ratios of *lengths* to other ratios of *lengths*.

As noted, Euclid builds his theory of proportion on this first proposition, vindicating his use of Definition V.5. The application of Proposition VI.1 to similar triangles begins immediately with VI.2, which states:

"If a straight line be drawn parallel to one of the sides of a triangle, it will cut the sides of the triangle proportionally; and, if the sides of the triangle be cut proportionally, the line joining the points of section will be parallel to the remaining side of the triangle."[39]

Euclid draws some auxiliary lines and then proceeds to relate the parts of each side of his original triangle to corresponding areas of various triangles within the resulting figure. In this, Euclid applies VI.1. He equates the ratios of the divisions on each side of the triangle to a ratio of areas within the triangle. He can establish that the two ratios of areas are equal. Therefore the ratio of the divisions of one side of the triangle equals the ratio of the divisions on the other. (I omit, as tangential, discussion of the converse, i.e., the second half of the proposition.)

Figure 19 should clarify:

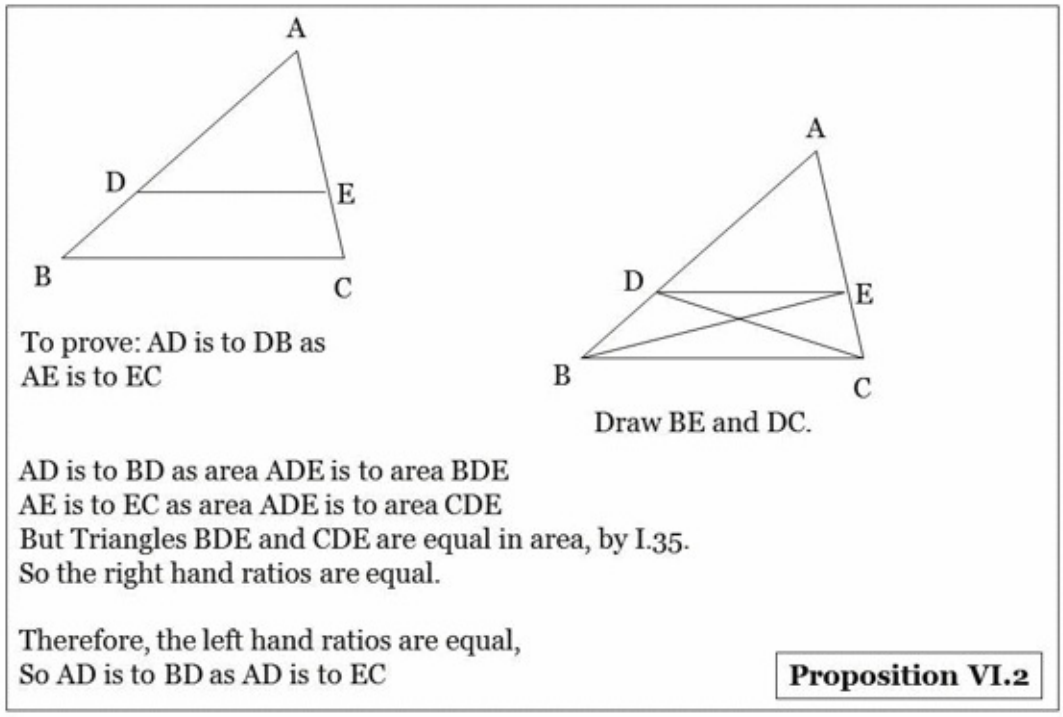


Figure 19

And here is the payoff: the proportionality of similar triangles (triangles with the same corresponding angles, which Euclid calls “equiangular”). Proposition VI.4 states:

“In equiangular triangles the sides about the equal angles are proportional, and those are corresponding sides which subtend the equal angles.”[40]

The proof is outlined in the Figure 20. It proceeds by placing the two triangles in a particularly helpful relationship and then drawing some additional lines to reduce the Proposition to VI.2. By now the pattern of abstract measurement should be familiar: Add, by construction, whatever features the diagram requires to render comparable the quantities being related. Then draw on previously known relationships to make the comparison. Euclid presents his measurement recipe and then helps you identify the relationship that you would, thereby, measure.

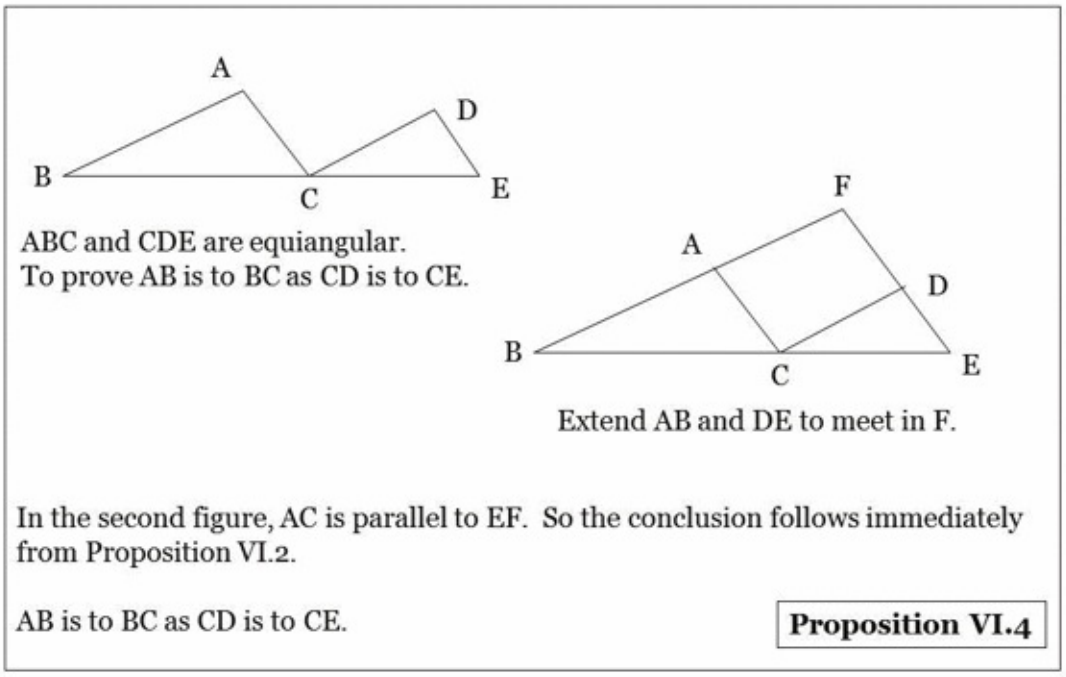


Figure 20

Now this is one of the vital underpinnings of indirect measurement. It is the foundation of an entire mathematical discipline known as trigonometry. It is the reason we can

- measure the locations of the planets, the stars and the galaxies
- understand the structure of the microscopic
- create blueprints for houses and skyscrapers
- create accurate maps of cities
- create plans for cars, boats, airplanes and a myriad of other industrial products
- design and build microscopic electronic parts, notably semiconductors, made of tiny pieces of silicon.

In short, it is one of the foundations of the modern world! As I have said, the Greeks put a much higher premium on constructability than we do today. It's as though, for them, to construct was to measure, to be constructible was to be measurable. They were not always successful, but their failures were regarded as problems that remained to be solved. For example, they were unable to trisect an angle or square the circle through straight edge and compass. But the challenge to carry out the Classical Greek program, by solving such problems, outlived both the collapse and eclipse of their civilization, outlived the very context that had made them important. Yet, even so, our understanding of mathematics, of

the constraints and needs of measurement, was vastly enriched as mathematicians ultimately discovered just why the Greek program could not be carried to completion.

The Greeks could not trisect an angle with straight edge and compass. But the theory of proportion gave them the means to trisect a line segment; indeed to divide a line segment into any prescribed number of equal parts. This is the meaning of VI.9, which says:

"From a given straight line to cut off a prescribed part."
The proof is outlined in Figure 21:

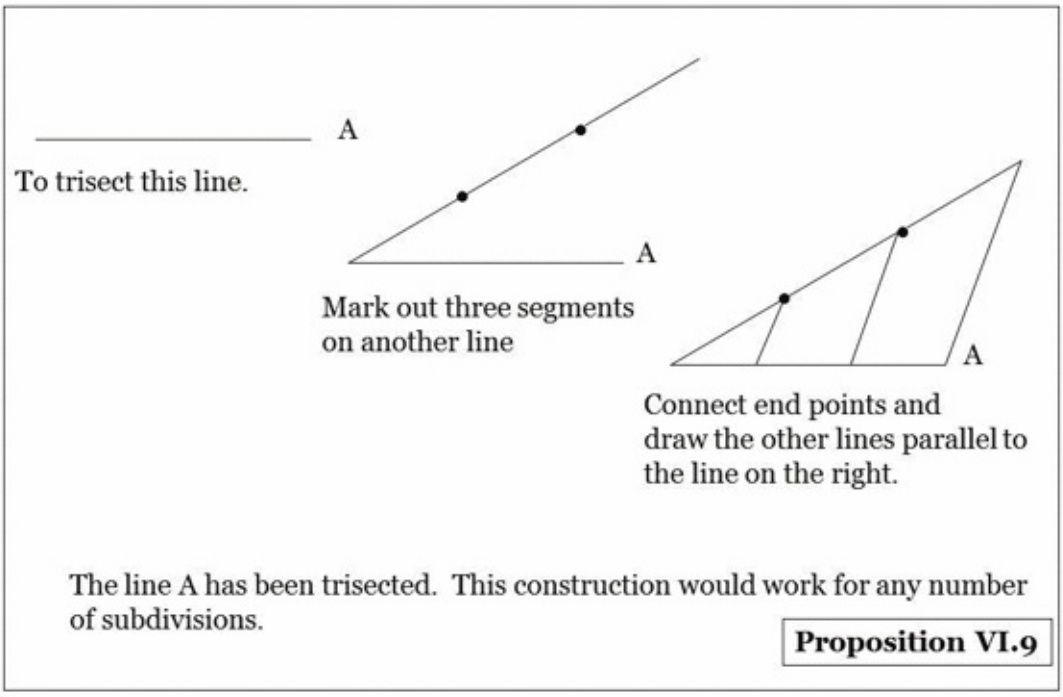


Figure 21

Euclid never offers a *formula* for area. Not in so many words. Even Archimedes does not; his "formula" for the area of a circle is expressed as an equality of the circle's area to that of a right triangle with a prescribed height (the radius of the circle) and width (the circumference of the circle).[41]

But, in asking for a formula, one speaks from the modern perspective. Euclid was not looking for a formula; he was looking for a standard way to express an area in geometric terms. For Euclid, to measure a magnitude was to *relate* a magnitude to another magnitude. It was not, *per se*, to attach a number even though, from a modern perspective, it creates a foundation to attach a number.

Euclid's entire line of development concerning area, from its beginnings in Book I, is directed towards establishing the essential relationships. As we shall see presently, his crowning success of that endeavor was VI.14, which is the essential fact that underlies the validity of the modern formula.

Euclid did not have our formula, but he *did* grasp the magnitude that he was trying to measure, he knew what he needed to establish, he knew the form in which he wanted to express that measurement, and, one should presume, he understood the meaning of VI.14 and why it was so important.

Proposition VI.14 reads:

"In equal and equiangular parallelograms the sides about the equal angles are reciprocally proportional' and equiangular parallelograms in which the sides about the equal angles are reciprocally proportional are equal."[42]

The final implications of this statement are found in VI.16, which is a special case of Proposition VI.14, and in VI.17, which is a special case of VI.16. We defer further explication of VI.14 briefly until we have stated VI. 16.

But first, Figure 22 outlines the proof of VI.14. In Figure 22, the diagram can be completed with the parallelogram EF lining up with the others because AB and BC are equiangular, i.e., have corresponding angles equal:

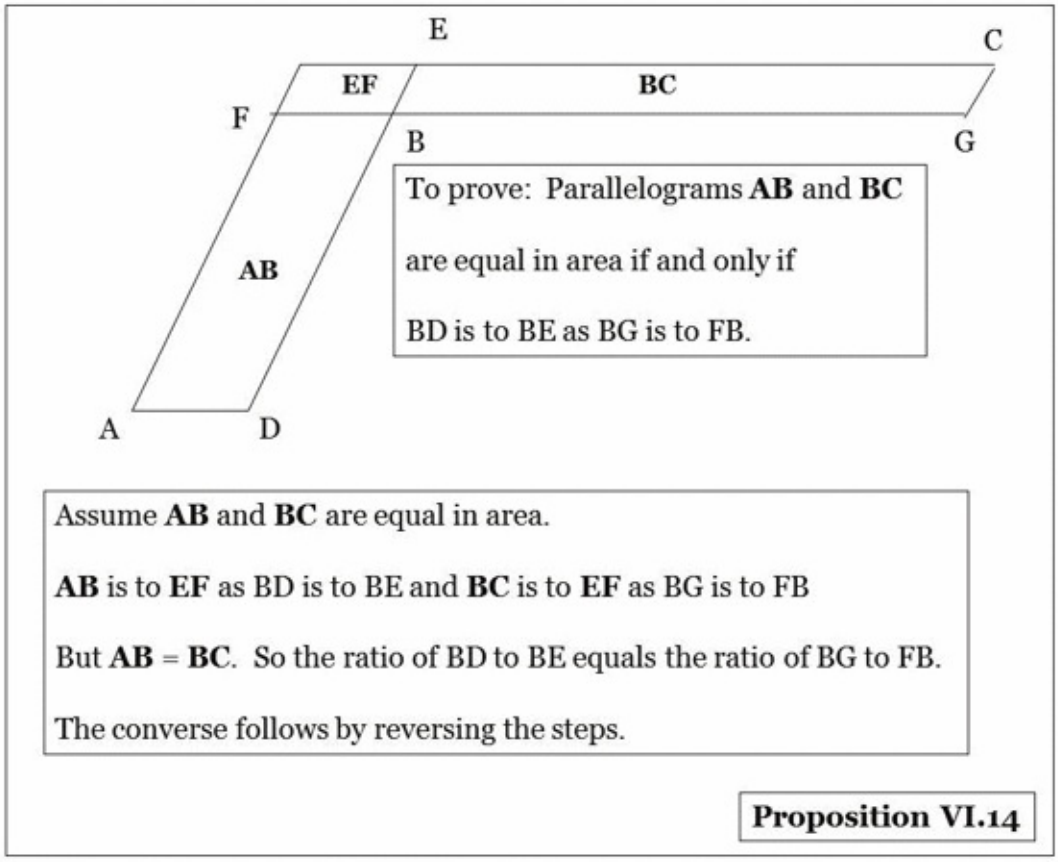


Figure 22

Proposition VI.14 is stated for general parallelograms. Proposition VI.16 is simply a specialization of VI.14 to rectangles and Proposition VI.17 is a specialization of VI.16 to squares. These propositions are stated, respectively, as:

“If four straight lines be proportional, the rectangle contained by the extremes is equal to the rectangle contained by the means; and if the rectangle contained by the extremes be equal to the rectangle contained by the means, the four straight lines will be proportional.”[43]

And

“If three straight lines be proportional, the rectangle contained by the extremes is equal to the square on the mean, and, if the rectangle contained by the extremes be equal to the square on the means the three straight lines will be proportional.”[44]

If one rectangle has sides A and B and a second has sides C and D. Euclid says

that their areas are equal precisely when A/C = D/B (in Euclid's language, A is to C as D is to B).

In Euclid's terms, A, B, C, and D are NOT numbers; they are magnitudes and Euclid *never* multiplied magnitudes. He does not express an area by a number but by a rectangle or, even better, by a square (a la VI.17), because there is a unique square having any particular area. For Euclid, with Propositions VI.14, 16 and 17 in the context of his earlier propositions, he has offered a *complete solution* to the problem of area for parallelograms and triangles.

But we moderns are used to numbers and never hesitate to invoke a unit of measurement. The modern instinct is to treat A, B, C, and D as numbers or, at least, of unknown numbers from the very beginning. One has a strong resistance to doing otherwise. One sees a letter, representing a variable, and, *thinking algebraically, one thinks of it as an unknown number*. So, from this perspective, one immediately concludes, from A/C = D/B that AB = CD (or A×B = C×D, to make the multiplication more explicit), where the lengths of the sides have been expressed in multiples of a chosen unit.

But, for Euclid, to say it one more time, these letters do not represent unknown *numbers*; they represent, in the modern sense, unknown *lengths* or, in the general case, unknown *geometric figures or quantities/magnitudes*.

Euclid has answered the question I posed earlier. *Because*, as we just saw, the areas of two rectangles are equal precisely when the respective products of their sides are equal, one can use the product, area = width times height, to measure the area of a rectangle. To put it another way: the modern definition of area *cannot be made arbitrarily*; it is only Euclid's *discovery* that makes the product a valid measure of area.

Euclid, as he pushed on to Propositions VI.16 and VI.17 understood the importance of these propositions and the need for them. But modern students, taught a pre-digested formula, do *not*. A full understanding requires both ways of looking at it: One needs Euclid's perspective, but one also needs the modern approach of counting squares.

I believe that this point is difficult to grasp even when it is pointed out, because it forces one to reflect on the origins of one's knowledge, to trace them back to our basic observations of the world. But if understanding means, as it must, grasping mathematical truths as truths about the world one inhabits, one does not

grasping mathematical truths as truths about the world one inhabits, one does not fully understand even elementary mathematics until one understands it the way that Euclid did, as relationships that one discovers in the world.

There is nothing more dangerous than a pre-digested formulation. I say dangerous, because it can lead one to think that one understands something when, in fact, in the full sense of understanding, one does not. One cannot correct a problem that one does not know one has. In the discipline of mathematics, understanding the achievements of the Greek geometers is one of the ways that we possess to find such limitations and to correct them.

Conclusion

The power of geometry is the power of indirect measurement. Much of this power derives from the Parallel Postulate and it does so despite the relativistic corrections that one makes to account for the effects on gravity on light. By the discoveries of geometry, one understands the structure of the universe and of the microscopic because one can relate the large and the small to the scale of objects that one can move and touch. The concrete measurements that one makes and interprets every day are made possible by the abstract measurement of Euclid's *Elements*.

As I explained and illustrated in Chapter 1, every one of Euclid's Postulates, Common Notions, and Propositions embody abstract measurement The proofs and constructions of Euclid's Propositions provide the links in his own deductions and also in our interpretations of our own measurements. Euclid's constructions provide the recipes by which the measurements would apply in any particular concrete instance. To understand Euclid's Corpus, as abstract measurement, is to more deeply understand geometry and Euclid's vision of it. It is to understand the over-arching structure of his work. And it is to appreciate, and to better understand, the concepts that one may have thought one understood already.

It is, finally, to create a foundation and an approach for understanding more advanced mathematics whose difficulties are more readily apparent.

One's grasp of the large and the small depends on the application of one's geometric knowledge to scientific observation and experiment. Leaving out the caveats, the high points and cornerstones of indirect geometric measurement include:

include:

- Light travels in a straight line
- A triangle is completely determined by two sides and the angle between them. (Proposition I.4) or, alternatively, by the three sides (Proposition I.8), or, alternatively by two angles and the side between them (Proposition I.26).
- Triangles are scalable: If corresponding angles are equal, their corresponding sides are proportional. (Proposition VI.4)
- "A straight line falling on parallel straight lines makes alternate angles equal to one another, the exterior angle equal to the interior and opposite angle, and the interior angles on the same side equal to two right angles." (Proposition I.29)

Euclid's *Element*s is one of the keys that have unlocked the universe to human understanding.

[1] Euclid, *Elements*, edited with notes by Thomas L. Heath (New York: Dover Publications, 1956), See Heath's notes to Book 1, Postulate 5

[2] John Stillwell, *Sources of Hyperbolic Geometry*, American Mathematical Society, 1996, original papers, with introductions by editor, p 35-63. Jeremey Gray, *Worlds Out of Nothing A Course in the History of Geometry in the 19th Century*, Springer-Verlag, London 2007, especially Chapters 19 and 20 (p 203-232) and Chapter 25 (p273-289). Marvin Jay Greenberg, *Euclidean and Non-Euclidean Geometries Development and History*, New York, W. H. Freeman and Company, 1983, Chapter 6, "The Discovery of Non-Euclidean Geometry, '' p 177-222. Peter Pesic, *Beyond Geometry Classic Papers from Riemann to Einstein*, Dover Publications, Inc., Mineola, New York

[3] Greenberg, Chapter 7, "Independence of the Parallel Postulate," particularly p 241-242

[4] Immanuel Kant, Critique of Pure Reason, Translated by Smith, Norman Kemp, 1965, New York, St Martin's Press, B42 (in the standard numbering)

[5] Pesic, "Geometry and Experience" by Albert Einstein, p 148-149

[6] Ayn Rand, *Introduction to Objectivist Epistemology Expanded Second Edition*, April 1979, p 8 in the paperback edition, "The purpose of measurement is to expand the range of man's consciousness beyond the perceptual level … By establishing the relationship of feet to miles, he can grasp and know any distance on earth … "

[7] Rand, p 195

[8] Harry Binswanger, "Selected Topics in the Philosophy of Science", 1987, available from the Ayn Rand Bookstore (www.aynrandbookstore.com), Binswanger makes a similar point in his discussion of parallel lines

[9] Euclid implicitly relies on this uniqueness in his proof of Proposition I.16, which lays the foundation for his proof of Prop I.27. For further elaboration on this point, see Heath's notes to prop I.16.

[10] Euclid, Book I, Heath notes for Postulate 5 under the section "The direction theory." Heath attributes this approach to Leibniz and quotes Gauss and Dodgson in rebuttal. In particular, Gauss's point is the one I have made in this section: "If it [identity of direction] is recognized by the equality of the angles formed with one third straight line we do not yet know without an antecedent proof whether this same equality will

also be found in the angles formed with a fourth straight line."

[11] Greenberg, Chapter 6, p 177-222. Gray, Chapters 19 and 20, p 203-232

[12] Greenberg, Chapter 7, "Independence of the Parallel Postulate", particularly p 241-242

[13] Herodotus, *Histories*, 1972, London, Penguin Books, Book Two, p 109

[14] Euclid, Prop I.29

[15] Euclid, Prop I.27

[16] Euclid, Prop I.33

[17] Euclid, Prop I.34

[18] Euclid, Prop I.20

[19] Euclid, Heath note following Prop I.35

[20] Euclid, Prop I.35

[21] Euclid, Heath comments on Prop I.35

[22] Euclid, Prop I.36

[23] Euclid, Prop I.38

[24] Euclid, Prop I.41

[25] Euclid, Prop I.43

[26] Euclid, Prop I.47

[27] Euclid, Prop II.13

[28] Euclid, Prop II.14

[29] Archimedes, *The Works of Archimedes*, Cambridge University Press 1897, "Measurement of a Circle", Proposition I equates the area of a circle to that of a right triangle with height equal to the radius of the circle and base equal to its circumference.

[30] Euclid, Definition VII.1, "A *unit* is that by virtue of which each of the things that exist is called one."

[31] Euclid, Prop VII.2

[32] Euclid, Definition I.23. "*Parallel straight lines* are straight lines which, being in the same plane and being produced indefinitely in both directions, do not meet one another In either direction."

[33] Sir Thomas Heath, A Manual of Greek Mathematics, Dover Publications 1963, in a section entitled "The irrational", Heath notes that the discovery, if made known, "would immediately be seen to throw doubt on so much of the Pythagorean proofs of theorems in geometry as rested on their (arithmetical) theory of proportion." (p 105)

[34] Euclid, Prop VI.1

[35] Euclid, according to Heath's notes following Prop X.5, Euclid's problem is that he uses a different definition for equality of ratios for numbers (Book VII, Definition VII.20) than he does for magnitudes (Definition V.5). Proposition X.5 expresses a relationship between a ratio of magnitudes and a ratio of numbers, but Euclid did nothing to bridge the gap between these two definitions. Euclid treats numbers as magnitudes, but he cannot, by the same token, treat commensurable magnitudes as if they were necessarily numbers, as he does in his argument. However, as Heath presents, there is,within Euclid's framework, a remedy to this lapse in Euclid's argument.

[36] Richard Dedekind, *Essays on the Theory of Numbers*, "Continuity and Irrational Numbers", section IV "Creation of Irrational Numbers", Dover Publications 1963 from a 1901 English translation, German publication 1872

[37] Euclid, Definition V.5
[38] Euclid, Prop VI.1
[39] Euclid, Prop VI.2
[40] Euclid, Prop VI.4
[41] Archimedes, “Measurement of a Circle”
[42] Euclid, Prop VI.14
[43] Euclid, Prop VI.16
[44] Euclid, Prop VI.17

Chapter 4

Numbers as a System of Measurements

Euclidean Geometry is about shapes that exist in the world, viewed from an abstract perspective. Shapes can be measured and their measurements, the relationships of their parts to appropriate standards, can be expressed by numbers. Numbers, however, do not exist, as such, in the world. Rather, they are means of expressing measurements, means of expressing or specifying a relationship to a unit. Numbers are a way of looking at the world, of identifying a particular quantitative relationship. Triangles are objects in the world, viewed from a particular perspective. Triangles are a kind of *thing*; numbers are a way of *looking at* things, of comparing things.

In Chapter 2, I focused on magnitudes and I discussed numbers as *measurements* of magnitudes. My interest in *this* chapter is to understand numbers as a *system* of measurements, measurements of a particular kind.

Numbers are a system by virtue of their interrelationships. But mathematical relationships among numbers reflect quantitative

relationships among the things that they measure. A number measures the relationship of a quantity (a multitude or a magnitude) to a unit or to another instance of the same kind of quantity.[1]

The sum of two multitudes is the composite of the two multitudes, the multitudes being taken together as one multitude. The sum of two numbers measuring these multitudes is the numerical relationship of the composite to a unit. As for the sum of two magnitudes, see Chapter 2.

To anticipate later developments in this chapter, everything that we know about numbers reduces ultimately to counting units. Real numbers apply to magnitudes. Their sums derive from counting. Three feet plus four feet is seven feet. One counts the unit, namely the foot. If one subdivides, one counts a smaller unit. Three fourths of a foot plus two fourths of a foot is five fourths of a foot. Three units plus two units is five units, the unit being one fourth of a foot.[2]

The relationships that irrational numbers name are quantified by comparing irrational numbers to rational numbers. One places the irrational number between two rational numbers, sufficiently close together, and, in the appropriate context, one regards the irrational number as indistinguishable, in application, from either of the rational numbers.

One's appreciation of natural numbers begins when one learns to count and to relate the numbers that one reaches to the multitudes that they measure. Adding one unit to a collection increases their number to the next number in the sequence. One grasps that numbers are related to each other by succession and realizes that one can reach any particular multitude by counting high enough. Practical limits aside, no matter how high one counts, one can always count higher.

That one can reach any multitude by counting high enough means that there is no actual infinity. All multiplicity is finite. An infinite number would have to be one that could not be reached by counting.

As I discussed in Chapter 2, Aristotle recognizes this point when he states the so-called "Axiom of Archimedes" during his arguments against the existence of an actual infinity. He writes: "…for every finite magnitude is exhausted by means of any determinate quantity however small."[3]

One can always count higher because the concept of number is open-ended. One need not commit to any particular number as the highest one that anyone will ever need. As Ayn Rand put it, "A concept is like an arithmetical sequence of *specifically defined units*,

… including *all* units of that particular kind."[4] Concepts, including the concept of number, are open-ended.

The domain of positive integers, also known as natural

numbers, consists precisely of the first number, namely one, and

any successor of one. These are not facts that one deduces; they are

identifications. They are something that one *grasps* about the nature of multiplicity and about our most basic means of measuring

it.

The value of identifying the

domain of positive integers

consists in knowing that one knows how to measure multitudes,

because the number one and its successors suffice to name any

multitude that one will ever need to measure.

One identifies that domain of numbers by first grasping

some particular numbers and then by grasping the relationships

between the numbers in the domain, the primary one being

succession.

Notice that the

scope of the concept of number (i.e., of the

natural numbers), once grasped, is independent of one's context.

The scope of the concept does not change when one discovers a

need for particular numbers that one hasn't previously encountered.[5] Typically,

one does not even notice that the new number is new, for that number is already subsumed within the

number

system, as such. One does not name each number

individually. Rather, the decimal number notation provides a

system for designating any number that one will ever encounter.

One grasps that system as a totality, as a graspable domain,

because numbers vary along a single dimension that is isolated

when one learns to count and when one understands what the

counting is accomplishing.

Over the eons since men learned to count, the concept of

number has been expanded a number of times, notably to rational

numbers, negative numbers, and real numbers. These extensions

have larger domains than that of natural number. But the domain

of the *original concept of number*, now distinguished from the wider concept by the alternative designations “natural numbers” or

“positive integers”, remains the same.

I said that one isolates the domain of natural number when

one learns to count. This isolation is essentially ostensive. Like

ostensive identifications generally, it provides a foundation for

further discoveries.

For example, one identifies the domain of integers by

specifying its relationship to the domain of natural numbers. In Chapter 2, I indicated how negative numbers arise as they apply to magnitudes. They arise, as well, as differences between multitudes. If one considers a larger multitude in relation to a smaller, the difference is positive; if one considers the smaller in relation to the larger, the difference is negative. A multitude, as such, can only be positive. But a *difference* of compared multitudes can be negative.

My immediate interest, though, is to specify the *domain* of integers. First, include zero. Second, include the negative of any natural number. This prescription is a complete characterization of the domain: Any integer is a natural number, zero, or the negative of a natural number.

The same approach applies to identifying the rational number domain, as well. Any rational number is a ratio of integers. (Integers are included as special ratios with a denominator of one.) This, however, does not quite characterize the domain. One must realize three things. First, one cannot divide by zero; there are no rational numbers like 6/0. Second, two distinct pairs of integers can represent equal ratios. For example, 6/9 is the same ratio as 8/12. One way to compare them is to reduce each to lowest terms. Since

both 6/9 and 8/12 are equal to 2/3, they are equal to each other.

Finally, one needs to address the fact that, say, 2/(-3) = (-2)/3.

Taking that into account, since every rational number can be

uniquely reduced to lowest terms, one can now specify the domain

of rational numbers as consisting of all valid ratios of integers

expressed in lowest terms with a positive denominator.

In pattern, one identifies the domains of integers and

rational numbers by first identifying how these numbers relate to

natural numbers and then, on that basis, establishing their domains

in relation to the domain of natural numbers.

In specifying the domain of rational numbers, I have drawn

upon their relationships to natural numbers. But those

relationships are multifaceted and I have not done justice to them.

First, most fundamentally, a rational number involves

subdivision of a unit. One divides a foot into 12 equal parts and says

that each part, of one inch each, is one twelfth of one foot. And this

means simply that twelve times this smaller unit is equal to the

original unit. In this way, a fraction reduces to counting, indeed it *is* counting, as viewed from the other side of the relationship. From

the side of inches, one says that twelve inches equals one foot. But

this *same relationship* is expressed, from the other direction, by saying that one inch is one twelfth (1/12) of a foot. One's new

fractional perspective, then, is simply a new way of looking at the older, counting, perspective.

Because an element of a subdivision is just a unit viewed from the perspective of a larger unit, the arithmetic of natural numbers extends to the arithmetic of fractions. Five inches is a multiple of one inch just as five marbles is a multiple of one marble. The situation is no different if one writes 5 inches as 5/12 feet. An inch – 1/12 of a foot – is a unit like any other, so one can also add multiples of the fraction 1/12. Thus 5/12 + 5/12 = 10/12 and 12 times 5/12 = 5.

In one sense, one can regard the arithmetic of fractions as an extension of the arithmetic of natural numbers. But, in a deeper sense, the arithmetic concepts have not changed. Rather, they have been applied to a new context with a different basic unit. And that basic unit, in any specific case depends on the specific fraction under consideration. The arithmetic of fractions is really an *application* of the arithmetic of natural numbers.

Indeed, all relationships of fractions follow from the relationship of the smaller unit, represented by the denominator, to the larger unit. Thus, one establishes that 4/12 = 5/15 by showing that they bear the same relationship to one – that they are both one

third of one. Or, alternatively, one shows that they are both the same multiple of 1/60, namely 20/60.

Finally, it is important that fractions can be compared as to size, both to natural numbers and to each other. Just as 3 is larger than 2, 3/7 is larger than 2/7. And because any fractional unit can be further subdivided, one can compare fractions of different denominators. For example, 3/5 is greater than 4/7. This can be seen by expressing each as a multiple of a still smaller unit, namely 1/35. Thus, 3/5 = 21/35, while 4/7 = 20/35.

When I specified the domain of rational numbers, my discussion took such relationships for granted. But to understand the rational numbers, one needs to understand their relationship to the world. And one key to such understanding consists in identifying the fundamental and derivative ways that fractions relate to natural numbers, as catalogued above.

As noted, one does not prove that the domain of natural numbers is what it is. One simply grasps its domain in the process of forming the concepts of particular numbers and grasping the way that these numbers all relate to each other.

The case of integers and rational numbers are similar in key respects. One discovers them and forms their concepts by

identifying quantitative relationships in the world, such as the relationship of a part to a whole (fractions), that these numbers can be used to measure. But notice, once again, that these new kinds of numbers are expressed by means of natural numbers. This further reflects the cognitive need I have just mentioned: Understanding of, say, rational numbers requires relating rational numbers to whole numbers. With the domain of natural numbers as a base, one establishes the domains of integers and rational numbers by reducing and relating these cases to the domain that one already knows: the domain of natural numbers.

A number is a type of measurement that is used to measure multitudes and magnitudes, that names the relationship of a multitude or magnitude to another multitude or magnitude and in particular that names the relationship of a multitude to an individual or the relationship of a magnitude to a standard. A number stands for a relationship, a quantitative *relationship*, not a quantity. A number *measures* a quantity by *identifying* the relationship of the quantity to the applicable unit. A magnitude is a type of *quantity*; a number is a type of *measurement*.[6]

The principle of measurement omission applies to numbers in the same way that it applies to triangles. In the case of a triangle, one focuses on the lengths of the edges and the magnitude of the

one focuses on the lengths of the edges and the magnitude of the angles, ignoring their color and the immaterial imperfections in the edges and angles. In the case of number, one focuses on the relationship to the unit and omits consideration of the type of multitude or magnitude to which it is applied.

The relationship to a unit is a relationship among real entities and their attributes. And relationships among numbers, generally, conceptualize relationships among real entities and their attributes. But these numerical *relationships* do not depend upon which particular multitude or magnitude, or to what type of magnitude, they might be applied.

In the case of natural numbers, the specific numerical *values* of the numbers are treated mathematically. But the nature of the particular *units* being counted is not treated mathematically.

For the laws of arithmetic do not depend on the specific units being counted.

In treating triangles mathematically, one discovers (or learns) trigonometry to relate the sizes of the angles to the lengths of the edges. In treating numbers mathematically, one learns addition and multiplication. In the realm of numbers, indirect measurement begins with arithmetic.

Although the decimal system helps one to deal with

numbers, one's identification of the domain of multiplicity does not depend upon having found a way to express every number. The ancient Greeks did not have such a system, but they grasped the concept, both of number and of its domain. Archimedes, in one of his extant works, developed a scheme to estimate the number of grains of sand that would be required to fill up the universe up to the fixed stars.[7] In this very pursuit it is clear that he would have invented notations for more numbers if he'd seen a need for it. His calculations that established his limits already reflected the fact that anything beyond was lacking only a name, a name that would be supplied whenever the need arose. There was, however, no mystery in how these new numbers would relate to the numbers that had already been named.

The concept of number is not limited to the particular numbers that one has named or that one has so far encountered.

Measurement of Continuous Magnitudes

As I discussed in Chapter 2, irrational numbers arise in the measurement of continuous magnitudes. I indicated, in that connection, the fundamental importance of the axiom of Archimedes, and our consequent ability to employ decimal expansions to achieve any required precision.

But if one can always find a suitable approximation, why

does one need irrational numbers?
Since irrational numbers arise in the measurement of

continuous quantities, of magnitudes, I begin by taking a closer

look at the measurement of magnitudes.
To begin with, systematic direct measurement of

magnitudes requires subdividing a standard. For example, to find

an answer within millimeters, each meter must be divided evenly

into a thousand subdivisions. Each measurement with this

precision is expressed in millimeters, in thousandths of a meter.

Expressed in terms of meters, then, the measurement necessarily

results in a rational number.
However, this measurement outcome, a rational number, is

due to the *method* of measuring, *not* to the specific characteristics of the *magnitude* being measured. Whatever the nature of these magnitudes, whatever the actual magnitude might be, this *method* guarantees that rational numbers will be used to express the result.

In other words, this is a

methodological distinction, not a

metaphysical one.
To illustrate this point, notice that there is an alternative to

this methodology, at least in some cases. For example, one can

measure out by a geometric construction with straight edge and

compass, a length of $\sqrt{2}$, an irrational number. Whether or not

Euclid's constructive approach can provide a systematic approach to measurement, it is, inherently, no more and no less precise than

laying out multiples of a subdivided unit.
But there is a more fundamental point to consider,

stemming from the fact that any measurement of magnitudes is valid within a particular, finite level of precision. As Ayn Rand put it,

"But more than that, isn't there a very simple

solution to the problem of accuracy? Which is

this: Let us say that you cannot go into infinity, but

in the finite you can always be absolutely precise

simply by saying, for instance: 'Its length is no less

than one millimeter and no more than two millimeters.'"[8]

A measurement is usually specified by a number, but an

explicit identification of its precision requires or entails placing the

result within a *range* of numbers. However, a finite range of numbers cannot distinguish a rational number from an irrational

number. Direct measurement of a magnitude cannot distinguish a

rational number from an irrational number. In other words, in the

context of direct measurement of magnitudes, the need for

irrational numbers does not come up.

Then how can it come up at all? All measurement, even

indirect measurement, ultimately resolves into a series of direct

measurements. So how can the need arise for the broader category

of measurement when no such need arises for the concrete direct

measurements that constitute, in toto, the substance of any specific

measurement?

The problem, at least to some extent, is with the question.

The real issue regarding numbers is not with the nature of

measurement, but with the way that one designates the results. Do

numbers, as such, designate *specific relationships*?

It is certainly true that, in any

application to a specific

context and in relation to a specific quantity, a measurement

determines a range of possible numerical values. Any identification

of a concrete *quantitative* relationship is subject to contextual precision.

But numbers are the conceptual form that the identification

takes; they are the means of identification, of *specifying* the relation of a magnitude to a standard. Specifications, *in general*, are applied contextually, but the relationship that a specification

names is specific, independent of any application, with its related precision requirements. If a *number* did not stand for a precise relationship then *precision-intervals* would also have no precise meaning, which would mean that even approximations would be

meaningless. Numbers are required to specify approximations and

they can only do so if they name precise relationships.

For example, if one says that a measurement is 3 inches

plus or minus one eighth of an inch, one is, clearly, saying that the length cannot be 3 3/16 inches. If one says that 3 inches is one's best estimate, one, thereby, makes a precise identification of one's best estimate. One is naming a specific number to express that estimate. One does not say that the *length* is exactly 3 inches, but one *does* say that one's *estimate* is exactly three inches. The number 3 has an exact meaning, independent of the precision

context of the measurement, of the quantitative relationship, to

which it is applied. The *number* 3 is not subject to contextual precision intervals; numbers are the *means of expressing* precision intervals.

If one also says that one's measurement could be off by as

much as 1/8 of an inch, this statement puts precise limits on one's

measurement. One is leaving open the possibility of an error of

more than 1/16 inches, but one, quite as clearly, denies the

possibility of an error as much as 3/16 inches. Indeed, one is

denying that the length could be as much as one trillionth of an

inch greater than 3 1/8.

Every number involved in this example has a precise

meaning; each names a specific mathematical relationship to a unit.

The number 3 represents the best estimate of length, as it relates to

the unit of length. The fraction 1/8 is the largest possible deviation

and the number, 3 3/16, represents a number outside that range.

That numbers designate precise relationships is easily seen

directly with the application of whole numbers to measurement and

is fairly easily seen with rational numbers, as well. First, when one

counts, one relates a multiplicity to a unit. One can make counting errors, but five is five, whether one is counting objects or meters.

The case of rational numbers, as applied to magnitudes, reduces to counting. As applied to a magnitude, such as length, the denominator of a fraction specifies a subdivision of a larger unit and a fraction is simply a multiple of that smaller unit. An inch, for example, is 1/12 of a foot. A length of 5/12 feet is a number of inches, namely 5 inches. One's physical identification of an inch as a subdivision of a foot is, certainly, subject to precision limits, but one finds 5/12 of a foot by counting inches. To count iterations of inches is to count iterations of these smaller units.

Whatever precision limits may apply to subdivision of a foot into 12 inches, these limits apply to the side of subdividing, not in the counting. In regards to feet, 5 feet is an exact count of the intervals one takes to be feet; similarly, 5/12 of a foot is an exact count of what one takes to be inches. The relationship, the ratio, of 5 feet to one foot is a *quantitative* relationship, a relationship between two quantities, subject, in one's identifications, to precision limits. The relationship of 5 to 1 is a *mathematical* relationship, relating the *counts*. The relationship of 5 to 1 is independent of the units one is counting and, certainly, independent of any precision limits regarding the comparability of those units. As I've already pointed out, specifications of the

specific objects being counted are omitted, are an omitted measurement.

So far, whole numbers and rational numbers designate *determinate* relationships to a standard. But, as identifiers of potential *quantitative* relationships, they are also *distinguishable*.

Indeed, any two numerical relationships of a magnitude to a standard are potentially distinguishable at some finite level of precision. And we saw in chapter 2 that, as applied to quantitative relationships to a unit of magnitude, any two different magnitudes can be distinguished by a rational multiple of that unit lying between them. Numbers designate *specific* mathematical relationships; any two different numbers potentially identify *distinguishable* quantitative relationships.

We have, just now, directly observed the specificity of natural numbers and of rational numbers. But these do not exhaust the relationships that a magnitude can have to a standard. There remain irrational numbers, designating relationships to a standard, or relationships between two magnitudes, such as the relationship of the side of a square to its diagonal, that cannot be expressed by rational numbers.

The difference between rational numbers and irrational numbers does not represent a metaphysical distinction among

magnitudes: the difference between rational and irrational numbers consists in the *means* of measurement, in how such numbers are specified, not with the *object* of measurement. And it depends upon one's choice of standard. For example, if the length

of the side of a square is one inch, then the length of the diagonal,

in inches, is the square root of two, an irrational number. But if the

diagonal were the standard, the situation would be reversed. If, say,

the diagonal were an inch, the irrational number would be

represented by the side, rather than the diagonal.

But what does it mean, for example, to say that the ratio of

a square's diagonal to a side is √2? It means two things. First,

leaving aside physical limits, it means that a sufficiently precise

squarewouldexhibitaratio arbitrarilycloseto √2.Secondly,thisis

the case for no other number. Any rival approximation, different

from √2, would, at some point, fall outside the precision range.

From a geometric perspective, that is, from a universal perspective,

the ratio of the diagonal to the side is √2. To say that a square has

no material imperfection is, among other things, to say that the

ratio of its diagonal to a side does not differ materially from √2.

Irrational numbers are no more and no less precise than

rational numbers. Qua measurements of quantitative relationships,

measurements by both rational numbers and irrational numbers are subject to the same precision limit. Qua numbers, precision is

an omitted measurement; both rational numbers and irrational numbers designate specific, distinguishable, mathematical relationships. As Mathematical relationships, both reduce to counting, directly so for rational numbers and indirectly for irrational numbers, through their relationships to rational numbers.

The important mathematical difference between rational numbers and irrational numbers consists in how they are specified, in that it is much easier to specify a rational number than an irrational number. To specify a rational number, two whole numbers, a numerator and a denominator, suffice. But, to specify an irrational number one specifies its relationship to rational numbers and to specify this relationship is significantly more complex than specifying the relationship of a rational number to a whole number.

But the essential method is the same: One specifies an irrational number by specifying its mathematical relationship to other numbers that have already been specified. Such specifications arise in the context of indirect measurement and can take many forms. The principal burden of the rest of this section is to show how that is done.

how that is done.

That numbers designate specific relationships is vital because indirect measurement is vital. Indirect measurement, as such, requires an ability to condense a series of mathematical relationships into a single composite relationship. For example, one reasons, through a series of steps, that if $2x + 5 = 17$, then $x = 6$. To traverse these steps is to solve the equation. These steps constitute a mathematical argument, identifying a series of quantitative relationships that, taken in sequence, imply the result. The validity of the argument requires that each step designate a precise mathematical relationship.

In this example, one's first step is to subtract 5 from both sides of the equation. And, in performing this step, one presupposes that, "5 is 5." One presupposes that a specific number, distinct and distinguishable from all other numbers, is being subtracted from both sides of the equation. And this presupposition applies to the unknowns, x in this example, as well, unknowns that may, for general polynomials, turn out to be irrational numbers. Without this presupposition, that every real number is distinct and distinguishable from all others, mathematical arguments would prove nothing; mathematical argument would be impossible.

One first encounters irrational numbers when one moves

beyond direct measurement and beyond the required distinctions

of any one particular case. To support indirect measurement, it is

necessary to establish mathematical relationships that link other mathematical relationships. In this process, over and over, one encounters numerical relationships that cannot be resolved into the

ratio of two integers; relationships that cannot be expressed by

rational numbers.

As I have said, irrational numbers are specified by their

relationships to rational numbers and as I will catalogue shortly,

these specifications can take numerous forms. But to specify a

number is not to know everything about it. Consider, for example,

the equation, $2x + 5 = 17$. This linear algebraic equation *specifies* a unique number, namely the value of x that satisfies the equation.

One finds the numerical value of that number, x, by solving the

equation, by applying a series of mathematical relationships to

simplify the equation. But the key point is that such a process is

necessary. Without going through these steps, one would not know

that $x = 6$.

The same principle applies to more complex cases such as

the ratio of the circumference of a circle to its diameter or the ratio

of the diagonal of a square to one of its sides. One specifies

determinate relationships in each case, but one does not know,

automatically, that the first ratio is between 3 and 4 or that the

automatically, that the first ratio is between 3 and 4 or that the second ratio is an irrational number.

In the case of irrational numbers, the most important thing one wants to know is how big is it? How does it compare in size to other numbers, specifically to rational numbers? Which rational numbers are larger than it and which are smaller? For example, one finds that the ratio of the diameter of a square to a side is between 1 and 2. One narrows it down further by establishing that the ratio is greater than 1.4, but less than 1.5. And, of course, one can go still further. In these determinations, one is not limited to any prespecified level of precision, but any specific determination of relative magnitude involves selecting a finite interval, bounded by two rational numbers bracketing the irrational number. Any comparison to rational numbers requires the choice of some particular rational numbers. One chooses *some* level of precision adequate to the demands of any particular case. In any concrete context, some finite degree of precision is required; any required finite degree of precision is possible.

Notwithstanding the complexities of approximation, an irrational number represents a specific number, distinguishable from any other. Accordingly, in any general mathematical analysis, in any chain of abstract mathematical measurements, one carries the irrational number through the entire chain until the end. Only

when the analysis is finally applied to a concrete, with a specific

precision requirement, does one find a decimal approximation satisfying the required precision.

One cannot, directly, simultaneously, and explicitly, specify

the relationship of an irrational number to all rational numbers all at once. But one can do so indirectly, by specifying a *system* of successive approximations. One *approximates* because, in any concrete application to a quantitative relationship, one never needs,

or can achieve, infinite precision. A sufficiently precise

approximation, in the concrete, will always serve. However, one

creates a *system* of successive approximations because there is no *prior* limit to the precision that might be required, nor any limit to the precision that is mathematically achievable.

Consider how the approximation of an irrational number

works out in practice. Take the square root of 2 (in the usual

notation,√2).To specifythesquarerootof2 as the unique positive

number with a square equal two, is already to specify a *precise* number bearing a precise relationship to 2 and, therefore, a precise

relationship to 1. However, to establish √2 as a number, more is

needed. One must also establish it as a potential measurement of a

magnitude, as potentially naming a quantitative relationship. And

to do that, one must be able to compare it along the appropriate

dimension. One must be able to compare it, as to relative size, with

other measures of magnitudes, specifically, with rational numbers.

And, in that connection, one notes that a positive rational number, such as 3, with a square greater than 2 is too big, while a positive rational number, such as .5, with a square less than 2 is too small. Since the square of a positive number increases with its size, the positive value of √2 must lie between any two such numbers.

To specify √2 in relation to the number, 2, is easy. But to specify its *place among the rational numbers* requires more elaborate means. For example, there is a standard procedure for extracting square roots that is somewhat akin to long division. The algorithm generates an ever-lengthening decimal expansion of √2.

Now suppose, in some context, that one requires accuracy to six decimal places. Suppose that nothing, *in that context*, beyond six decimal places has any relevance. In such a case, limited in application to that context, there is *no meaningful distinction* between 6-place accuracy and 19-place accuracy. Decimals beyond the 6^{th} decimal point, "accurate" or not, are simply meaningless. In that context, the expansion, to six decimal places, *is* √2.[9] In the same sense, an expansion to 19 decimal places, being indistinguishable from the shorter expansion, is, *in that same context, also* √2. In such a context, if only six figures matter, any number in the series beyond six-place accuracy will have the required relationship to 2.

Yet, the sequence of approximations also offers further refinements, as needed, for more demanding contexts. These

refinements do not contradict the less demanding contexts because

the later approximations are only meaningfully distinguishable

from earlier approximations in the more demanding contexts. Sixplace accuracy is √2 in the first context, but not in the second; the place accuracy is √2 in the first context, but not in the second; the

placeapproximationis√2in *both* contexts. Conversely, the two approximations are

relevantly distinguishable in the second

relevantly distinguishable in the second

place accuracy whenever such accuracy is needed, but that 19-place

accuracy is only relevant or meaningful, only relevantly different

from less precise approximations, when it is actually required.

In the less demanding context,

all expansions that are

accurate in the first 6 decimals, *mean the same thing* and equally qualify as √2. In particular, the decimal expansion of √2 to six

decimal places *is*√2,as long as one holds the particular context. In sum, the *root-extraction*

process for generating a decimal

expansion of √2, or any other similar process,is a *method* to specify √2,amethodthatsupplies,foranyrequiredprecisionlevel,avalid

approximation to √2. Such a process guarantees that any specific

precision requirement that need be met can be met. The *process* sufficesto

specify √2 in any context that could ever arise. The

algorithm, as such, *simultaneously addresses* every potential finite precision requirement that could ever be demanded, serves to

distinguish √2,asrequired, fromanyother number,andtospecify

its ordering relationship (larger or smaller) with respect to any

rational number. To specify the *process*, is to *uniquely specify* √2.

Theseapproximationsto√2donotreflectanyimprecision,

lack of specificity of √2, or any inability to distinguish √2 from any

other number. The need for the sequence of approximations is not

to *specify* √2; it is to relate its magnitude to the rational numbers.

Considered as a number, the square root of two is distinct from any

other number. Were this not the case, the entire root-extraction

would be meaningless; to approximate √2 presupposes a specific

number that one seeks to approximate.

So when I say that all numbers within a certain range are

effectively equal, I am not, strictly speaking, saying that they are

equal qua numbers. They are, so to speak, *effectively* equal and that word, *effectively*, presupposes a precision context in which every number within the range of valid approximations bears the

required relationship to 2. In the unqualified sense of number,

there is no range because, as far as the concept of number is

concerned, the precision standard is an omitted measurement.

Every number, as such, names a distinct mathematical relationship

Every number, as such, names a distinct mathematical relationship,

is distinguishable from any other number.

But I

am saying something else. I am saying that *one* way to specify a number is to provide a *process*, such as the root extraction process, a *method* for determining suitable

approximations and that applies, universally, to achieve any

required degree of precision. The square root extraction process, in

exactly this sense, is a specification of √2.

Notice that I have now offered two different

characterizations of the square root of 2. In the first, I characterized

√2 in relation to the number 2. The number √2 is that number

whose square is 2. In the second case, I characterized it as the limit

of a root-extraction process.

The first characterization addresses the question, “Why do

I care?” I care because of the relationship of √2 to 2. The second

characterization addresses a different question. Namely, “What is

thenumerical valueof√2, as it compares with other numbers?” Its

numerical value is the limit of the root-extraction sequence of

decimal expansions.

These two characterizations provide two different

perspectives on the same mathematical relationship. Both

perspectives are important and meaningful. But each is directed towards a different question, addressing a distinct cognitive need.

Relating different ways of looking at the same thing has tremendous importance in mathematics, and, indeed, in all human investigations. So it is important to grasp the respects in which these two characterizations are the same, but also, the respects in which they differ. They are the same in that they have the same *object*; they pertain to the same number, the same mathematical relationship. They are different in that they involve two different *perspectives* on that relationship.

In a this-worldly sense, what do these two characterizations mean? And why do they characterize the same number?

As Ayn Rand points out, "… a word has no meaning other than that of the concept it symbolizes, and the meaning of a concept consists of its units."[10] So what units are embraced by my two characterizations of √2? What are the specific referents of these two characterizations?

Processes providing successive approximations to irrational numbers pertain to a higher level of abstraction than concrete measurements. They pertain directly to mathematical relationships among numbers, and only derivatively to quantitative relationships among magnitudes. To specify a concrete quantitative relationship, a single approximation will always suffice. To specify a

mathematical relationship, one must account for all potential precision requirements.
But the same basic principle applies: On the level of concretes, including the application of numbers to concretes, there

is always a standard of precision. One is looking, in relation to the

first characterization, for a number whose *square* is within, say, one ten-thousandths of 2. And one is looking, in the second

characterization (following the squareroot-expansion process) for

a *decimal expansion* out to, say, seven decimal places. In the first case, within the required precision, any number whose square is

within one ten-thousandths of 2 *is* √2. In the second case, and, again, within the required precision, any decimal expansion of at

least seven decimal places *is* the limit of the sequence.
I have, at this point, claimed to provide two distinct

characterizations of √2. But the question remains: *Why* do they characterize the same number?
It is important to realize that real work is required to

answer that question. One does need to establish a *mathematical* relationship between the two characterizations. In the first case, a

mathematician characterizes √2 in relation to its *square*, that square being 2. In the second case, one characterizes the number in

relation to other nearby numbers: to the decimal expansion of √2.
In pattern, a mathematician shows, for example, that if one

considers an expansion of √2 to 100 decimal places, the square of

that decimal expansion will approximate 2 within, say, 98 decimal

places. In other words, one *shows* that one can always get a square sufficiently close to 2 by finding enough places in the decimal

expansion of √2. And, by the terms of the first characterization of

√2, that decimal expansion, therefore, *is* √2: That is, the square of the decimal expansion meets the precision requirement.
There is nothing special about the precision requirement

(98 decimal places) I have placed on the square of the squareroot

extraction. The process works for *any* required precision level. The specific precision requirement does not affect the outcome, is

irrelevant to the final result. If the specifications agree in *any* precision context, then they agree in *all*. Accordingly, one can treat the specific precision requirement as an omitted measurement and

conclude, as mathematicians put it, that the limit of the expansion

is √2, which is to say that its square is 2. The two characterizations

of √2 do, indeed, characterize the same number.
Examples of irrational numbers are legion. The square root

of any positive rational number, expressed in lowest terms, is

irrational unless both numerator and denominator happen to be

perfect squares. Thus the square root of 9/25 is rational, but the

square root of 8/25 is not. In the same way, the cube root of any

rational number expressed in lowest terms is irrational unless both numerator and denominator happen to be perfect cubes. So the

cube root of 1/27 equals 1/3, a rational number, whereas the cube

root of 1/26 is irrational because there is no whole number with a

cube of 26. Similar statements apply to fourth roots, fifth roots and,

in general, to nth roots. The roots of most polynomial equations

with integer coefficients, such as $x^2 + x - 1 = 0$, are irrational.

Finally, the relationships between the angles of a triangle and the

lengths of its sides is almost always expressed by irrational

numbers, no matter what unit one chooses to measure length and no matter what unit one chooses to measure the angles. Within the

sphere of mathematical relationships, ratios of integers are the

exception rather than the rule.

It is important, also, to realize that one can *identify* or

specify a number without using a standard naming system to name it or without even naming it at all. As we saw with √2, the *meaning* of a number, the *scope of its referents*, does not depend upon, any particular *means* by which one *identifies* it.

Here, there is almost a continuum of possibilities, between

giving a specific number a name and specifying it indirectly. For

example, the ratio of the circumference of a circle to its diameter is

so important that it has been given a name, namely the Greek letter

π, also spelled out as “pi”. As a second example, there is a system

for designating roots. To illustrate, the square root of 2 is written

√2. This is not quite a name, but it is a standard designation. The

fifth root of three, written $3^{1/5}$ is a similar case. Finally, to relate angles to lengths, one can write $\cos 37^0$ to designate the ratio of a leg of a right triangle to the hypotenuse when the angle between them

is 37^0.

In all of these examples, one indirectly specifies a

numerical measurement by specifying its relationships to numbers

that have already been defined. The basic principle: One reduces or relates the newly known or the unknown to the already known.

Convergent Series and Sequences

There is another, very general, method of identifying irrational numbers and finding rational approximations to any required degree of precision. The ability to find rational approximations is important: An irrational number may be *specified* in any of the ways I have been discussing, but ultimately, to apply it directly to measuring magnitudes, one needs to be able to approximate it by rational numbers, most typically, today, by decimal expansions.

That general method for dealing with irrational numbers is provided by convergent series and sequences.[11] As a first example, remembering that the irrational number π is the well-known ratio between the circumference of a circle and its diameter, Leibniz was able to establish that

$$\pi/4 = 1 - 1/3 + 1/5 - 1/7 + 1/9 \ldots .$$[12]

As this series, converging to an irrational number, already illustrates, one cannot formulate a consistent mathematical theory of approximation without dealing with irrational numbers.

This example is a

convergent series. In general, a *series*, sometimes called an *infinite series*, is an infinite *sum* of specifically defined numbers, evaluated from left to right. The "partial sums"

are the subtotals one reaches along the way. For example, the

second partial sum is 1 – 1/3 = 2/3. The third is 1 – 1/3 + 1/5 =

11/15. These successive *partial sums* form an infinite *sequence*. A series is *convergent* when the

sequence of partial sums

is

convergent, i.e., can be shown to have a limit, to serve as successive

approximations, sufficient to meet any finite precision requirement,

to some definite number. My earlier example of a sequence of

decimal approximations to √2 is a convergent sequence.

The rule that determines this particular series, for π/4, is

manifest from the first few terms. One finds an approximation to

π/4 by simply summing a sufficient number of terms, starting from

the left, to reach the desired precision.

Now there are other series for π, equally elegant, that

converge much more quickly, but this one does converge. In this

particular case, no matter what precision might be required, a

mathematician can specify the number of terms required to achieve

and guarantee such precision. Indeed, a mathematician can write

down a formula to express the number of required terms as a

down a formula to express the number of required terms as a

function of the sought-for precision. Yet, to establish that a series,

or a sequence, converges, it is not always necessary to produce such

a formula.

Thataseriesconvergingtoafractionofπcouldfollowsuch

a simple rule is surprising, though hardly accidental. However,

defining a *rule* or a *formula* that determines each term in an infinite series is not the only approach that can both specify and

provide a sequence of converging approximations of irrational

numbers. Defining specific algorithms can be equally effective.

For example, a computer might apply the following

algorithm to calculate the square root of a number: Suppose that

the number is 8 and the goal is to approximate √8. Start with half

of 8, namely four. That's the first estimate. Now divide 8 by 4 to get

2 and find the average between 4 and 2. The average is three, your

second estimate. Proceed in the same fashion until the differences

between successive estimates become sufficiently small. When a

term *x* in the sequence is approximately equal to 8/x (the first step out of the two that would generate the next term in the sequence), it

follows that x^2 is approximately equal to 8, which is to say that x is approximately equal to √8.

In comparing this process to the series that I provided to

evaluate π/4,one findsone keyelementthattheyhave in common.

Both of them define a *process* to generate rational approximations, of any required accuracy, to a particular number.

Notice that in order to demonstrate that a sequence

converges to π one must have already specified this particular number, π, in some other way. Indeed, π was originally specified by reference to an external relationship, namely the ratio of a circle's

circumference to its diameter. This *specification* was, in fact, an *identification* of a particular geometric relationship. However, as I have already observed, to *identify* or *specify* a mathematical relationship is not to know everything about it. To establish, for

example that π is greater than three but less than four is a

discovery, adding to our knowledge about this ratio. But this determination is just that, a *discovery* about a specific number, identified and referenced in advance.

A number arises in a

context; it provides a numerical

measure of a relationship referring, ultimately, to the external

world, to a particular class of quantitative relationships. When a

mathematician finds a particular sequence of rational numbers to

approximate a particular number that has arisen in some context,

he is adding to his knowledge of a number that has *already been specified*.

In general, any

given irrational number can be expressed as

the limit of a sequence of rational numbers and can be

approximated to any desired accuracy by such a sequence.[13] Although

estimation can be difficult in particular cases, such a sequence can easily be *specified*, as follows. To simplify the discussion, it is enough to assume, and I will assume, that the given

number is positive:

The first term in the sequence is the largest integer less

than or equal to the given number. The second term is the largest

decimal expansion with one figure past the decimal point that is

less than or equal to the given number, Term number N is the

largest decimal expansion containing N – 1 figures past the decimal

point that is less than or equal to the given number.

Depending on

how a particular given number is *given*, it may, indeed, be very difficult to *explicitly* identify the terms in this sequence. Even finding the very first term, the largest integer less

than or equal to the number, may be very difficult. Yet if the

number is *given*, no matter how, the sequence that I've described is also given, is *indirectly specified*, as well.

Suppose that one is asked: What about numbers that are

never *given* in some way? There are, the argument continues, by Cantor's reckoning, an in-denumerable infinity of real numbers, of

which only finitely many will be encountered in anyone's lifetime.[14]

But, within a realitybased approach, this is not an issue.

One does not

construct quantitative relationships; one finds

methods to *measure* them. To isolate a *system* of measurements is to establish a method to measure whatever relationships, of a

particular type, one encounters or might encounter.

This relationship of a system of measurements to

quantitative relationships does not exclude hypothetical references

to *potential* quantitative relationships. Indeed, my very discussion was hypothetical: given a number, however specified, there exists a

corresponding converging sequence of numbers approximating the

given number to any required degree of precision. It exists in the

sense that it is specifiable in the fashion I have just illustrated.

Contrary to Cantor,[15] a system of measurements is not a completed infinity. To isolate a *system* of measurements is not to imply existence of every measurement that, if it did exist, would be

included in the system. That is not the purpose, nor the

achievement, of a system of measurements. Rather, it is to identify

a place for such measurements, to provide a *method* of dealing with any measurement of that type that does exist, that ever arises, in

any form in which it might arise.

To specify a

method of measurement is not to, somehow,

bring into existence, *even as a thought*, every particular measurement that might, one day, be applied to a concrete quantity

or geometric relationship. On the contrary, to provide for the

possible existence of a quantitative relationship is to recognize, first

of all, that quantitative relationships *of that type* exist. It is to recognize, secondly, that a particular quantitative *dimension* has been identified to which it would belong if it did exist. It is to

provide a *comprehensive method* to deal with any measurement of the type, to *find and distinguish* a place for any particular measurement along the particular quantitative dimension. It is to

provide for a *contingency*.

A contingency is not,

per se, a potential, nor is a

potentiality an actuality.
So a number can be specified by a sequence. What if one

starts from the other direction? What if one *starts* with a sequence?

After all, one of the ways that numbers can arise is as limits to

converging sequences of rational numbers.
Just any kind of sequence simply won't do; not all

sequences converge. But there is a particular kind of sequence that

will *always* converge. At least, it will always converge as long as there is *something for it to converge to.*
But is there, in fact, something for it to converge to? What

is the ontological status of the limit of the sequence? How, if at all,

does it relate to the world?
This was the

mathematical issue that confronted

mathematicians, such as Dedekind and Cantor during the second half of the nineteenth century. In a conventional formulation of the

question: Do all sequences that *should* converge have something to converge *to*?

The approaches and the answers that Dedekind and Cantor, at about the same time, provided are both taken as standard answers today. Both answers are accepted as uncontroversial. I will outline and discuss both approaches, and

both answers, to this question, at the end of the chapter. For now,

keep in mind that this question was taken seriously and, in my

view, should be taken seriously. It has a history and it is important

to understand that history.
But my immediate objective is to offer a this-worldly

answer to that question.
The kind of sequence I have in mind, the sequence that

should converge, as long as there is something for it to converge to,

is called a *Cauchy sequence*. A *Cauchy sequence* (a_1, a_2, a_3, … a_n, …) is, first of all, a sequence of numbers whose *i*th term can, for this discussion, be designated by a_i, (where i = 1, … ∞). The sequence is a *Cauchy* sequence whenever, given any ε > 0, there exists a positive number η such that whenever n > η and m > η, then

$| a_m – a_n| < \varepsilon$.[16]

For example, the sequence

½, ¾, 7/8, …, $2^{n-1}/2^n$

is a Cauchy sequence. Beyond a certain point in a Cauchy sequence, any two terms differ by less than ε, no matter how small a positive

number ε may have been chosen to be.

One point of defining such a sequence is that one doesn’t

need to know the *limit* of a Cauchy sequence to know that it *is* a Cauchy

sequence. A Cauchy sequence will always converge,

providing that one grants that the unique number that it converges

to actually exists.[17]

In the standard mathematical terminology, to say that all

Cauchy sequences of rational numbers (or of real numbers)

converge is to say that the real number system is *complete*.

For a realist account, my discussion of √2 already

illustrates the principle involved. A Cauchy sequence, as such, *as a sequence*, is not a number. But it *specifies* a number in exactly the same way that √2 specifies a number: It provides a rule for defining

successive approximations to any required precision, just as square

root extraction provides successive approximations for √2.

Indeed, a Cauchy sequence

specifies a number in just the

following sense. Whatever level of precision may be required, in any

concrete case, there is a point in the Cauchy sequence from which

every subsequent term in the sequence meets that precision requirement. Pick *any* one of those subsequent terms.

That chosen term, whatever it may be,

is the limit of the Cauchy sequence *within the required precision level*.[18] Just as I found for √2, the chosen term is *indistinguishable* from any "better" approximation that may occur later in the sequence.

Within the standard of precision they are all equal.

A Cauchy sequence indirectly specifies a mathematical relationship in exactly the same way that the algorithm for √2 does. Like the algorithm for √2, the Cauchy sequence of approximations offers further refinements, as needed, for more demanding contexts. These refinements do not contradict the less demanding contexts because the later approximations are only meaningfully distinguishable from earlier approximations in the more demanding contexts. The accuracy it provides, whenever such accuracy is needed, is only relevant or meaningful when that degree of accuracy is actually required.

In just this way, the Cauchy sequence, as a *system* of approximations, simultaneously provides *the limit* in *every* concrete instance according to the standard of precision in each case. From a realitybased perspective, this is what it means, the only thing it *can* mean, for a Cauchy sequence to converge. As applied to concretes, the sequence simultaneously supplies a valid approximation for every precision standard.

To say this slightly more technically: At any precision level, i.e., for any given $\varepsilon > 0$, one can choose *any* term a_n, for which $n > \eta$, to satisfy the precision requirement expressed by $\varepsilon > 0$. But the fact that the specific decimal expansions may depend upon which

term one chooses from the sequence is irrelevant here. It's

irrelevant because the term chosen in any particular context is

indistinguishable, *in that context*, from later terms in the sequence that might be required for a more demanding context. That a value

chosen in these more demanding contexts might be more precise *is only relevant and meaningful in those more demanding contexts*.

The Cauchy sequence, then, applies to

all cases, to all finite

levels of precision. It applies universally, measuring the designated

relationship in every single case. A number identifies a relationship

and a Cauchy sequence *acts* as a number by universally identifying a particular relationship. The limit is the relationship that it

identifies.

Now suppose, as given, two Cauchy sequences with general

terms a_i and b_i such that the related Cauchy sequence with general term $a_i - b_i$ has a limit of zero. Then the two Cauchy sequences have the same limit. For any required precision level, one can find a

point in both sequences at which all successive terms in each

sequence will simultaneously provide a numerical value to serve as a limit, a limit within the required degree of precision.

Unambiguously, the two Cauchy sequences identify the same

mathematical relationship. Conversely, any other Cauchy sequence

that converges to the same limit as the first sequence will be related

to it in the same way. (I here apply an insight of Cantor, who,

however, gave it a far different interpretation, as we shall see.)[19]

In particular, every decimal expansion is a Cauchy sequence. And, for every Cauchy sequence, there is a decimal expansion that converges to the same number as the Cauchy sequence. For example, to find the first 100 decimals of the expansion, simply find a point in the Cauchy sequence from which all subsequent terms agree to 100 decimal places.[20] Then, since these terms will all agree to 100 decimal places, any one of these terms provide the first 100 decimals of the decimal expansion. One can apply this procedure, in turn, to every required term in a decimal expansion and generate, in this way, a decimal expansion with the same limit as the Cauchy sequence.[21]

In just this fashion, the use of infinite series provides a general mathematical approach to successive approximation. But, because Cauchy sequences can converge to irrational numbers, one cannot develop a theory of approximation that doesn't include irrational numbers as potential limits of a sequence of approximations.

A theory of approximation is also needed to justify our approach to direct measurement. A theory of approximation, appealing to the Axiom of Archimedes, discussed in Chapter 3, is needed to establish that one really can approximate magnitudes, to

any required accuracy, by means of rational numbers.

But, returning to an earlier point, rational numbers, like

irrational numbers, characterize the *means* of measurement; the distinction between rational numbers and irrational numbers is

not, as I have observed, a metaphysical distinction pertaining to

magnitudes. It is rather, a *mathematical* distinction pertaining to the *means of identification* of quantitative relationships. And, as a mathematical abstraction, to deny irrational numbers is, to deny

the mathematical theory of approximation, and, ultimately, to deny

the possibility of measurement of continuous magnitudes.

At the end of Chapter 1, I discussed the need for precision

in mathematics. What I said there applies here, as well.

Fundamentally, the need for irrational numbers arises from a

difference between mathematics and engineering. In engineering,

with a concrete application in mind, one can always identify the required precision in advance. But the application of mathematics

is open-ended. It applies to all engineering problems that will ever

be tackled and all levels of precision that will ever be required.

There is simply no way to anticipate the level of precision that may,

someday for some reason, be required by someone.

By techniques that necessarily involve irrational numbers,

mathematicians are able to achieve what an engineer cannot. A

mathematician can analyze complex chains of mathematical

relationships without ever losing precision. In this way, mathematics can provide any required level of precision without having to know, in advance, the requirements for any concrete case. By contrast, any prior approximation by rational numbers would fail at some finer level of precision that may one day be required. The validity and universality of mathematical conclusions depends on the ability to analyze complex chains of mathematical relationships without ever losing precision.

One never encounters irrational numbers when one makes concrete numerical measurements. But irrational numbers are utterly required at the next level of abstraction to establish mathematical relationships ranging from the most mundane to the most important. And, ultimately, they are required to ground the approximations involved in all concrete measurements of magnitude.

To summarize, a Cauchy sequence provides unlimited *mathematical* precision by providing a *system* of approximations, a system that presupposes *some* finite precision requirement in any particular context, but one that is adequate to *any* finite precision requirement, irrespective of context.[22]

So what is the status of the concept of irrational numbers? As Ayn Rand once put it, in answer to a question about imaginary numbers:

“If you have a use which you can apply to actual reality, but they do not correspond to any actual numbers, it is clearly a concept pertaining to method. It is an epistemological device to establish certain relationships.”[23]

As Ayn Rand characterizes “concepts of method”:

“Concepts of method designate systematic courses of action devised by men for the purpose of achieving certain goals.”[24]

She elaborates, as follows:

“The concepts of method are the link to the vast and complex category of concepts that represent integrations of existential concepts with concepts of consciousness, a category that includes most of the concepts pertaining to man’s actions. Concepts of this category have no direct referents on the perceptual level of awareness (though they include perceptual components) and can neither be formed nor grasped without a long antecedent chain of concepts.”[25]

Based upon the account of irrational numbers that I have provided, irrational numbers, and, specifically the use of convergent sequences to identify them, are a concept of method, a

methodological device to keep track of distinguishable mathematical relationships. The specific “course of action” (or *one* of them) involved in this case can be characterized as: Providing a series of successive approximations to measure magnitudes to any required accuracy. I have shown how irrational numbers relate to reality, but this is also where consciousness comes in. Irrational numbers arise in an abstract setting as relationships between magnitudes that cannot be specified as ratios between integers (rational numbers), but *can* be specified by relating or comparing them to rational numbers. The context provided by this abstract setting is essential to their meaning, to the particular way that an irrational number relates to reality.

Taken together, the rational numbers and irrational numbers are referred to as “real” numbers. Like rational numbers, real numbers can be compared: any real number is either larger or smaller than any other real number. Between any pair of real numbers there is at least one rational number and at least one irrational number. All the operations of addition, multiplication, etc., can be defined and applied to all real numbers. Any real number can be approximated to any desired degree by a sequence of rational numbers. In specifying the way that real numbers relate to rational numbers, one specifies its domain.

The Real Number Line

In Chapter 2, I discussed the real number line as a metaphor. But it is more than this. Considering both real numbers and an infinite straight line from an abstract perspective, real numbers can be put into correspondence with a straight line. When real numbers are identified with points on a line, considered geometrically, the result is called the “real number line”. It remains to examine the value and the conceptual validity of this correspondence.

One of the great unifications in mathematics was the development of analytic geometry, for which the decisive steps were taken, independently, by Descartes and Fermat in the early 17th century.[26] Analytic geometry integrates geometry and algebra by using a coordinate system to express geometric shapes. For example, if x is the horizontal axis and y is the vertical axis, the formula $y = 2x + 5$ represents a straight line consisting of all coordinate pairs (x,y) that satisfy the equation. Thus, (1,7) is on the line because $7 = 2 \times 1 + 5$. One says that the equation “represents” a straight line because the locus of points satisfying the equation is a straight line.

The essential underpinning of this setup is represented by the two axes, each of them assigning a number to every point of the

the two axes, each of them assigning a number to every point of the axis. On this foundation, a coordinate pair, based upon these axes, assigns a pair of numbers to each point in a plane.

Each axis is a copy of the real number line.

Now, even in the 17th century, the idea of associating numbers to points was not altogether new. Its ultimate genesis is the use of number to measure magnitude, which has its roots in antiquity. The Greeks used line segments to represent magnitudes and their understanding of the concept, magnitude, included length, area, volume, and, for Archimedes, weight or force.[27] To measure out a distance on a line segment is to measure a magnitude.

In the final analysis, to associate points on a line with numbers is to say that number can be used to measure magnitudes. Today we create tape measures marking out multiples and subdivisions of a standard length, such as a foot or a meter. The utility and validity of tape measures should not require an argument. Moreover, the fineness of the subdivisions in a tape measure is a physical limitation, not a mathematical one. We are limited by the available precision, the thickness of the markings, and our ability to discriminate.[28]

In sum, neither the validity nor the utility of assigning numbers to points on a line should be in dispute, nor is there any

trace of Platonism in using tape measures. The philosophical issue, as always, revolves around the mathematical treatment of precision.

In discussing triangles in Chapter 1, I defended the mathematical *treatment* of lines as totally straight, totally continuous, and without width. In this, I argued that the deviations from straightness, the lack of microscopic continuity, and the width of actual lines, is irrelevant in an appropriate mathematical context. Such deviations only become relevant in more concrete settings in which *something* is known about the required or available level of precision.

In discussing irrational numbers, I argued that irrational numbers are necessary to a mathematical theory of approximation. I argued, further, that they make it possible for mathematicians to integrate a complex chain of mathematical relationships without unnecessarily losing precision. I showed specifically how irrational numbers arise in the context of indirect measurement when passing from concrete measurements to abstract mathematical relationships. Irrational numbers are a concept of method, as is the derivative concept of real number.

The notion of a *real number line* pulls these concepts together. I've discussed all of the elements to this integration. But there is one significant step remaining: In assigning a

there is one significant step remaining. In assigning a

correspondence of *all* real numbers to points on a line, one passes beyond the ability to distinguish points on actual physical lines.

Now the first point one should notice is that there is really

nothing very new here beyond our previous discussions. In any

application of mathematics to a concrete, one necessarily deals with

the specifics of the concrete, such as whatever limits to precision

are inherent in the specific context. The application of real numbers

is a case in point. The results of the mathematics apply precisely to

any and all concretes, within the respective precision limits for each

concrete. That it does so is inherent in the approach mathematics

takes to approximation; it's inherent in the technical definitions of

limit that forms the foundation of its theory of approximation.

The application of numbers to a line is different from their

application to concrete measurement only insofar as the line is

treated as a geometric abstraction, as discussed in Chapter 1. But to

show how an irrational number corresponds to a point on an

abstract line *is* to show how it corresponds to a point on each concrete line covered by that abstraction.

After choosing a zero point and the location of the number

one, it goes like this. Within the limits of available precision, one

knows how to subdivide a concrete physical line: to find a

corresponding point on the line for a rational number. Any

irrational number can be specified as the limit of a convergent

sequence of rational numbers. Following Corvini's approach in her analysis of Zeno's paradox,[29] for any concrete line, and within any specific context, one locates the irrational number by simply

following the sequence of rational numbers until one can no longer

meaningfully distinguish the point one has reached from the

mathematical limit of the sequence. Similar to Corvini's analysis, at

that point, one has located the irrational number on the line.[30] It remains to observe that this process will apply to any physical

situation, whatever the precision available in and appropriate to

each context.

Having indicated how the irrational number measures *each*

line in the category, I have, thereby, specified how it applies to the

category. Since the same process works for every irrational number (and also for every rational number), I have also, thereby, specified

a correspondence between real numbers and the mathematical line.

The concept of this relationship i.e., the concept of the real

number line is a concept of method. It is a concept of method for

the same reasons that irrational numbers are concepts of method.

As with that case, by showing how it applies to each particular, I

have specified, as one must, how it relates to reality. But as with

irrational numbers, this specification involves and requires taking a

specific abstract perspective of each of the particulars.

This is the essence of the philosophical justification of the

approximation processes, of taking mathematical limits, on all

levels of mathematical abstraction.
This viewpoint also justifies the process by which irrational numbers were discovered in the first place. When one proves that the diagonal of a square is incommensurate with the sides, one is specifically dealing with the concept of a line from a mathematical perspective, precisely in the ways that I have discussed throughout. One is establishing precise mathematical relationships inherent in a certain kind of shape.
Such is the philosophical perspective: the account of how the mathematical concepts and methods correspond to the world. But it has built upon a specifically mathematical perspective, starting from the insights of the ancient Eudoxus to the more modern form that these insights take in the work of, Cauchy and, appropriately reinterpreted, of Dedekind and Cantor.
And this brings me to the modern approach to real numbers, the approach pioneered by Dedekind and Cantor.

Mathematical Rigor and Philosophy

Cauchy was the first to provide a rigorous definition of limit that could be successfully applied to the calculus. But then, as the story goes, the question became: How do we know that a sequence converges to *something*? What does it converge to?[31]

As Jeremy Gray puts it, “he [Cauchy] took it for granted

that if an increasing and a decreasing sequence of magnitudes ultimately differ by an arbitrarily small amount then they converge to a common limit."[32]

Cauchy's definitions of limits and his applications of those definitions ushered in a new standard of rigor in mathematical thought. But Cauchy's ultimate appeal, as I take it, was to the external world. One does not create the world; one studies it. One measures it. And so, Cauchy did not evidence worry, any more than Newton had, about such questions of ontology.

Yet there was an issue here, one that needed an answer.

And increasing the level of mathematical rigor was an important, and certainly a worthy, objective.

Unfortunately, the new insistence on rigor, and the new approach to rigor, that developed in mathematics originated most specifically in nineteenth century Germany. It developed in the Zeitgeist of the new Kantian philosophy and the currents of German Idealism and phenomenology that spawned in its wake. Standards, including standards of rigor, do not develop in a vacuum. One's views of what needs to be proven, of what constitutes a valid proof, of what it means for something to be true or valid, are *philosophical* issues.

German mathematicians belonged to many different

German mathematicians belonged to many different

philosophical camps and disagreed violently on very fundamental

issues.[33] But, generally speaking, they were looking to put mathematics on a more rigorous foundation and German

philosophy, especially phenomenology, provided the context and

theater of that endeavor and the debate surrounding that endeavor.

By the end of the nineteenth century, the following pattern

had emerged in treatments of key mathematical abstractions, first,

for example, in regards to irrational numbers, and, later, for positive integers.[34] At the risk of over-simplification: 1.

Start with a view, and sometimes a statement, of the basic

properties historically associated with the concept.
2.

Construct a conceptual *model*, focusing on form and

ignoring content.
3.

Prove, very rigorously, that the model has those basic

properties.
4.

Proclaim the new model to be the actual *content* of the

concept that one was trying to elucidate or make rigorous.

As Dedekind put it, for example, his goal, and his

achievement, was to “demand that arithmetic shall be developed

out of itself."[35]

In this process, one abandoned any connection to an external referent. Rigor, it turned out, had nothing to do with establishing the connection of a mathematical concept to the world. Indeed, mathematicians now considered reliance on the properties of magnitudes or on "geometric intuition" to be the primary source of vagueness, of lack of rigor in mathematics.

Rigor, in the new approach to mathematics had much to do with internal consistency, or internal logic, and nothing to do with external reality. And the basic role of a model was, sometimes explicitly, more to provide confidence in the *consistency* of a mathematical abstraction than to provide an object of mathematical thought, even a *constructed* object of mathematical thought.

So, for example, if two mathematicians came up with competing constructions, the model builders might argue about which one was better. And they might look for holes in the opposing proposal. But, failing that, there was no fundamental conflict between the two constructions. They would regard the two models as, simply, two different ways of achieving the same end. And if one could establish a cross-reference between the two constructions, one would consider them to be equivalent, to be ultimately

indistinguishable in any way that matters.[36]

An early example of the new approach to rigor was

Dedekind's theory of irrational numbers.

Dedekind and the Modern View of Number

Dedekind's answer to the question that Cauchy had left

unanswered started, *implicitly*, with the Axiom of Archimedes, that there is a rational number between any two real numbers. I say

implicitly. The version of real numbers that Dedekind constructed satisfies the Axiom of Archimedes. But since Dedekind was not

concerned to demonstrate a connection to any other notion of real

numbers, he did not appeal to that axiom in his *development*. Yet, insofar as Dedekind's development was *motivated* by the idea of separating a line into two segments by a point on the line, he relied

on the fact that any two points on the line can be separated by a

rational number (once a unit has been chosen.) In this sense, and only in this sense, Dedekind appealed to the Axiom of Archimedes.

As I discussed in Chapter 2, the Axiom of Archimedes

implies that any number can be distinguished from any other

number by finding a rational number that separates them. But

that's a realist perspective. Dedekind's explicit goal was to avoid the

use of geometry and, indeed, to avoid any appeal to magnitudes of

any kind that might exist in the world.[37] Dedekind's task was to create a new mathematical reality, working with the raw material of

the rational numbers.

So Dedekind offered a new view of number: A number is a

cut in the set of rational numbers. A cut is a bifurcation of the set of rational numbers into two sets, sets that one may designate as L

and R, such that every rational number is in one of the two sets and

every member of L is less than every member of R. (Think *left* for L

and *right* for R)

As an easy example, L might be the set of rational numbers

less than or equal than 9/5 and R would be the set of all numbers

greater than 9/5. In this case, 9/5 is the largest member of L. On

the other hand, one might, alternatively, choose 9/5 as the smallest

member of R instead of the largest member of L. As in this case,

every rational number produces two Dedekind cuts. Dedekind

himself pointed this out and answered that the two cuts represent

the same number, namely the rational number that produces the

cut.[38]

As a more interesting example, L might include, first, all

negative rational numbers and, in addition, all positive rational

numbers whose *square* is less than 2. R would include all positive rational numbers with a square *greater* than 2.

One

thinks of the first Dedekind cut as being the real

number 9/5 and of the other as being the square root of 2. But, in a

literal sense, and in the intended sense, these *cuts* in the rational numbers, these *divisions of the rational numbers* into two halves, *are* the real numbers. There is no other choice on Dedekind's terms, because Dedekind steadfastly avoids any external referent in

the definition of irrational numbers. There is no other sense in which these irrational numbers can be said to *exist*.[39]

To develop his theory, Dedekind proceeds to define

addition and other arithmetic operations of these Dedekind cuts.

For example, the *sum* of two Dedekind cuts 1 and 2, with left sets L_1 and L_2, is a Dedekind cut with a left set that I'll call *L*. To define L is to determine a Dedekind cut because the right set of the Dedekind

cut is simply the complement of the left set consisting of every

rational number to the right of all the rational numbers on the left.

One *defines* that left set *L* as precisely containing all distinct sums that result from adding a member of L_1 to a member of L_2. In this discussion I have used subscripts simply to distinguish the left sets

associated with two Dedekind cuts.

Any rational number corresponds to a Dedekind cut,

namely the cut *at* that rational number. Accordingly, one verifies that the arithmetic of those "rational" Dedekind cuts, those that are

defined by rational numbers, precisely agrees with the older

arithmetic of rational numbers. A mathematician creates symbols

to express such ideas, but I'll say it in words: The Dedekind *sum* of two Dedekind cuts corresponding, respectively to *two rational numbers*, is precisely the Dedekind *Cut* that corresponds to the *ordinary sum of those two rational*

numbers.[40]

To express this symbolically: Suppose that s and t are

rational numbers and that $L_{(s)}$ and $L_{(t)}$ are the left setsof the Dedekind cuts corresponding to s and t. (I use parentheses to

distinguish this use of subscripts form the earlier use as in L_1.)

Similarly, $L_{(s + t)}$ is the left set of the Dedekind cut of the rational number s + t. Then, by Dedekind's definition,

$$L_{(s)} + L_{(t)} = L_{(s + t)}$$

As a reminder, the left side of this equation consists in the

Left set consisting of all sums $x_1 + x_2$ where x_1 is in $L_{(s)}$ and x_2 is in $L_{(t)}$.

Next, one defines order simply: If L_1 is a subset of L_2, *and*, in addition, L_2 contains at least two rational numbers that are not in L_1, then one says that $L_1 < L_2$. Otherwise, if only the first condition holds, one says that $L_1 = L_2$.[41] In the first case, L_1 is less than L_2, because L_2 contains more than one rational number lying to the right of every rational number in L_1. One has to say that L_2 contains at least *two* rational numbers not in L_1, because of the technical issue noted earlier: In regards to any cut at a rational

number, that rational number can be assigned, indifferently, to

either set.

One also verifies various laws of arithmetic such as that

addition is independent of order (A + B = B + A), *etc*.[42]
The entire process, of course, presupposes and requires

that one already knows how to add and compare rational numbers.

So, in a typical Kantian twist, one starts with an antecedently

developed concept of the rational numbers, constructs something

out of these raw materials, and then changes, or ignores, the

original meaning of these raw materials to embed them into the

newly minted system of real numbers, that is, of Dedekind cuts.
Finally, it turns out that one can use a Cauchy sequence to

define a Dedekind cut, a Dedekind cut that one takes to be the *limit* of the Cauchy sequence. Namely, one defines *the left set, L, of the limit Dedekind cut* to include precisely any rational number that is

less than or equal to *all but a finite number of terms* in the Cauchy sequence. In this fashion, Dedekind has a solution to the

completeness question: What do Cauchy sequences converge to?

Properly defined, they converge to Dedekind cuts!
One more technical question remains. Now that one has

expanded the realm of numbers to include Dedekind cuts, what

happens if one tries to repeat this process? What happens if one

creates cuts in the expanded set of Dedekind cuts? Or, alternatively,

what happens if one looks for the limit of a Cauchy sequence of

Dedekind cuts with the Dedekind cuts playing the role that the

rational numbers play in Dedekind's construction? And it turns out

that a cut in the set of Dedekind cuts, or the limit of a Cauchy

sequence of Dedekind cuts, can always be *identified with* a

Dedekind cut.
Notice that I referred to an *expansion* of the realm of

numbers. And I spoke of i*dentifying* something constructed in one fashion with something else constructed in a different fashion. My

very expression betrays the modern point of view. The modern

view, to wit, is that *it doesn't matter* that rational numbers no longer refer to ratios of whole numbers. *It doesn't matter what things actually are or what they actually mean, just how they*

relate to each other.[43] In that view, unthinkable before Kant's phenomenology, it doesn't matter that *rational* numbers *no longer* refer to something external. Nor that *irrational* numbers have lost all external reference.
From the modern perspective, all that matters is the

relationships that these new numbers have to each other, the

relationships between the ideas. If Dedekind cuts at rational

numbers have the same arithmetic and ordinal relationships that

the rational numbers do, then, from the modern perspective, they

are indistinguishable from the rational numbers. What *does* matter is that one can use the rational numbers as bootstraps to *construct* the real numbers, as, in this case, bootstraps to construct Dedekind

cuts. And, this accomplished, it only remains to then embed the

rational numbers from which one began into the set of constructs,

of Dedekind cuts, that results.
Form triumphs over substance. Relationships among ideas

trump and replace any relationship to reality. Like Kant's

phenomenological world, the only world that matters, in this view,

is the one that we construct ourselves.
Prior to Kant, castles in the air started in the sky. But a

Kantian castle in the air generally *appears* to start on the ground.

Then, like an Indian rope trick, the connection to the ground is

removed after the castle has been built.
Dedekind knew exactly what he was doing.
As I explained in Chapter 2, Eudoxus's criterion for the

equality of two ratios is a complicated way of saying that any

irrational number can be approximated, to any required precision,

by rational numbers. By the Axiom of Archimedes, there is at least

one rational number between any two irrational numbers. Any two

numbers can be distinguished by a rational number lying between

them. So a particular irrational number is completely *determined* or *specified* by the set of rational numbers less than it.
But there is a world of difference between finding a way to

identify, specify, or distinguish an external relationship versus

taking an *identification* to be an *independent, non-referential, object*. When one, as a concept of method, uses numbers to identify

relationships in the world, one regards numbers as a *means of awareness*. But when one constructs numbers as a non-referential

object, one cuts all ties to the world and treats an idea in one's head

as a *self-sufficient object of awareness*, an object with no reference to anything external.
Dedekind cuts were *not* a *rediscovery* of the criterion of Eudoxus; Dedekind was quite aware of Euclid's definition in Book

V of his *Elements*. Indeed, Dedekind took pains to distinguish his definition of real numbers, as Dedekind cuts, from a prior

viewpoint, that some attributed to J. Bertrand, that "an irrational

number is defined by the specification of all rational numbers that

are less and all those that are greater than the number to be defined."[44] Dedekind considered *this* view "common property of all mathematicians who concerned themselves with the notion of the

irrational." He wound up:

"if, …, one regards the irrational number as the

ratio of two measurable quantities, then is this

manner of determining it already set forth in the

clearest possible way in the celebrated definition which Euclid gives of the equality of two ratios."[45]

Dedekind did not share this regard for ratios of measurable

quantities. He said that Bertrand's notion, "has no similarity

whatever to mine inasmuch as it resorts at once to the existence of a

measurable quantity, a notion which for reasons mentioned above I wholly reject."[46]

As the goal of his entire investigation, Dedekind sought a

new kind of rigor, a rigor that, by his lights, could not be grounded

by reference to actual magnitudes or relationships between

magnitudes existing in the world. He stated his intention, thus:

"For, the way in which irrational numbers are usually

introduced is based directly upon the conception of extensive

magnitudes, which itself is nowhere carefully defined, and explains

number as the result of measuring such a magnitude by another of

the same kind. Instead of this I demand that arithmetic shall be

developed out of itself."[47]

Dedekind understood the connection of his theory to

Eudoxus and he also understood what made it different. Eudoxus

had sought a way to measure ratios between magnitudes. But

Dedekind determined to ignore magnitudes altogether, to develop

arithmetic "out of itself", to "define irrational numbers by means of the rational numbers alone."[48] Number, said Dedekind, "are free creations of the human mind." Real numbers, for Dedekind were a

certain kind of object that one constructs out of the rational

numbers.

So when Dedekind characterized real numbers as being a

certain kind of splitting of the set of rational numbers into two

subsets, he *meant* it literally. When he "identified" the rational numbers with specific Dedekind cuts, a cut at the point determined

by the rational number, he betrayed an unconcern with what

rational numbers themselves measure. And when he defined

arithmetical operations, ordering relationships, and limiting

processes on these pairs of subsets, the all but explicit implication is

that the *only* thing that matters is that these constructions all relate to each other in a certain way. That they relate in the very way that

mathematicians had hitherto attributed to *actual* numbers, i.e., to numbers standing for quantitative relationships.

With Dedekind, number lost the referential character that

it had always had. Constructed objects, as such, became the objects

of mathematics. To the question, “What does a Cauchy sequence

converge to?” Dedekind answered, in effect, “It converges to a

constructed object with no referential content.”[49]

Dedekind was not alone. The end of the nineteenth century

saw numerous similar attempts to provide more rigorous accounts

of the real numbers. Indeed, Dedekind rushed into print with his

own account of real number to establish priority over competing efforts by Heine and Cantor.[50] In general, the various competing attempts, like Dedekind’s, involved constructions, shared the

objective of avoiding reference to geometry, and offered their

constructions as the actual referents of the concept of a *real number*.

Most prominent, besides Dedekind’s, were the two, very

similar, but separate proposals offered by Cantor and Heine, as well

as a somewhat different proposal of Weierstrass.[51] These attempts shared a common theme. They singled out, took their starting point

from, some method by which one might *distinguish* one real number from another. And then they elevated this *method* into the *object*, as the real number that it might have served, from a realist perspective, to have distinguished.

For example, Cantor’s idea (and Heine’s) was to

define a

real number as a Cauchy sequence of rational numbers. But, since

two different Cauchy sequences can converge to the same number,

one had to define a so-called *equivalence relation*. One compares two Cauchy sequences by subtracting one from the other: In this

way, one creates a new Cauchy sequence by subtracting the first

term of one from the first term of the second, the second term of

the first sequence from the second term of the second sequence,

and so on for all corresponding terms. If the new sequence

converges to zero, then the two compared sequences are regarded

as *equivalent*. From a technical mathematical perspective, one says that the two Cauchy sequences belong to the same *equivalence class*.

In symbols, if the first sequence has terms a_1, a_2, a_3, … a_n, …

and the second sequence has terms b_1, b_2, b_3, …, b_n, … then the two sequences belong to the same equivalence class if and only if the

sequence with terms $a_1 – b_1$, $a_2 – b_2$, $a_3 – b_3$, …, $a_n – b_n$, … converges to 0. A real number, for Cantor and for Heine is an *equivalence class of Cauchy sequences*.[52]

To the question, “What does a Cauchy sequence converge

to,” Cantor and Heine provide a very curious answer. Namely, a

Cauchy sequence converges to itself, to the equivalence class to

which it belongs.

The proposals of Weierstrass, Dedekind, Cantor, and Heine

were (and are) all regarded as satisfactory proposals. Dedekind’s

proposal became the most standard construction and the clumsier

construction of Weierstrass is no longer taught today. But, from a formal perspective, they are all regarded as equivalent.

For example, the proposals of Dedekind and Cantor are formally equivalent in the following sense: One can put Dedekind's real numbers into one-to-one correspondence with Cantor's real numbers, a correspondence that preserves all arithmetical and ordering relationships.

In one direction, corresponding to any Dedekind cut, one can construct a corresponding Cauchy sequence.
For this discussion, it is enough to assume that the Dedekind cut represents a positive number. Start by finding the largest *whole number* in the left set of the Dedekind cut. That's the first term of the Cauchy sequence. The second term in the Cauchy sequence is the largest decimal expansion that contains just *one numeral* to the right of the decimal point and is also contained in the left set of the Dedekind cut. For the third term, one finds the largest left-set decimal that contains no more than *two numerals* after the decimal point. Continue in this way indefinitely. The resulting sequence is a Cauchy sequence, that is to say, a representative of an equivalence class of Cauchy sequences, that is to say, a Cantorian or Heinian real number.
Conversely, to define a Dedekind cut from a Cantorian real number, pick a Cauchy sequence from the equivalence class and proceed as I indicated earlier: Namely, the corresponding Dedekind

proceed as I indicated earlier. Namely, the corresponding Dedekind

cut is the Dedekind limit of the Cauchy sequence. It is the Dedekind

cut for which *the left set, L,* includes precisely any rational number that is less than or equal to *all but a finite number of terms* in the Cauchy sequence. One verifies that this determination does not depend upon

which Cauchy sequence one selects from the equivalence class.

Thus, one checks that both of these relationships are well-defined

and that they establish a one-to-one correspondence between the

two domains. And one checks that the rules of arithmetic, of

ordering, and of limits correspond exactly.
From the modern perspective, then, these various

proposals are completely equivalent. Either alternative is equally

satisfactory. It really doesn't matter whether one thinks of real

numbers as Dedekind cuts or as equivalence classes of Cauchy

sequences. As formal systems they stand in one-to-one

correspondence and the *relationships between their elements* are identically preserved in that correspondence. There is no important

difference *because* the ontological status of real numbers, what real numbers actually *are* is unimportant. All that matters are the *relationships* between the elements.[53] It's important that one provide a construction of *some* kind, but the precise construction that one adopts simply doesn't matter. Real numbers, on this view,

are not the *means* of awareness; they are the *object* of awareness.

But even this object has only a formal significance.
The only thing that the modern viewpoint absolutely

forbids is imparting any *referential character* (that is, any reference to the external world) to the objects that one creates. For to do so would be to abandon the Cartesian certainty purchased by these constructions. The modern view avoids referential content *on principle*. That late nineteenth century mathematics started down this path was by no means inevitable. In the avowedly referential, realitybased account of numbers I have given in Chapter 2, as well as the current chapter, I have drawn freely on mathematical ideas that were either known in the late nineteenth century or, in some cases, *contributed* by the very mathematicians, notably Dedekind and Cantor, most responsible for the modern turn. From a mathematical perspective, no further erudition is required to pursue a realitybased approach to number.

There was no mathematical necessity for Dedekind and Cantor to lead mathematics down this path. But, particularly in Germany, there was a philosophical context that had paved the way, a context that had fashioned this particular path into an open highway.

Kant's phenomenology provided the most important underpinning. Kant held that the entire world of experience was the construction of our own senses and intellect. By contrast, the world that actually exists was held, by virtue of those very senses and intellect, to be unknowable in principle. As to content, the

intellect, to be unknowable in principle. As to content, the phenomenal world was a free creation of the human senses; but its form was constrained by the particular structure of our sensory apparatus and our intellect, as manifested, respectively, in *forms of intuition* and Kant's *categories*. That the universe appears orderly is not because it really is orderly. Per, Kant, we would have no way of knowing whether the universe is actually orderly one way or the other. On the contrary, our senses and intellect are solely responsible for all *appearance* of order in the world of experience.

The constructions of Dedekind and Cantor are faithful renderings of this basic outlook: Forget about external reality. Reality is too vague! It will only interfere with the rigor that mathematics now requires. Structure is all that matters. Consciousness is its own object. Accordingly, in typical Kantian fashion, the German mathematicians took the very tools by which earlier mathematicians had studied quantitative relationships, the criterion of equal ratio by Eudoxus and the Cauchy sequences of Cauchy, and made them the original objects of mathematics, standing in for the external relationships that they had been fashioned to study. Faithful to the impetus of Kant's *Critique of Pure Reason*, the *means* of awareness became the *object* of awareness.

The paradox is that mathematics, the unacknowledged science of measurement, the study of quantitative relationships, managed to survive. The disaster in the foundations of mathematics

and the Kantian trappings persist to this day. But so, in my view, does the study of quantitative relationship. I return to this question in Chapter 6.

In this regard, the final irony, which may, in part, shed light on how mathematics has survived, is this: At the same time that Dedekind, Heine, and Cantor were disastrously wrong, they were, in a different sense, almost right. As I say, they took tools designed to study quantitative relationships, tore them, by brute force, from their original context and enshrined them as *number* in place of the very relationships that they had been designed to measure. But, by the same token, both the criterion of Eudoxus and equivalence classes of Cauchy sequences do, in fact, distinguish particular mathematical relationships, in precisely the ways that I have discussed in Chapter 2 and throughout this chapter.

Officially, all ties to the world are broken. But I suspect that mathematicians know, implicitly, and on some level, what these formulations actually mean; and that they regard Dedekind's and Cantor's definitions as, simply, precise ways of getting at it. For, there is, in fact, a germ of truth here. One can, in fact, restore the discoveries of Dedekind, Heine, and Cantor to their rightful context, to a referential, realitybased context.

Dedekind did not invent the idea of using rational numbers to distinguish irrational numbers. He himself quotes a nineteenth

to distinguish irrational numbers. He himself quotes a nineteenth century expression of that idea and correctly identifies its historical origin in Eudoxus's criterion for equal ratio. And Cantor did not invent the practice of doing arithmetic on infinite series. Euler, for one, is famous for his flamboyant engagement in that very practice.

But while the prices that Dedekind and Cantor each tried to extract was too high, indeed gratuitous, they each contributed something important to a proper mathematical understanding.

Eudoxus gave us a way to *compare* ratios, but he never provided a way to *add* them. But Dedekind did. To the extent that one can specify two irrational numbers by specifying their respective relationships to rational numbers, Dedekind showed that one can, thereby, specify the sum of those irrational numbers in the same terms, can specify the relationship of the sum of the two irrational numbers to the rational numbers. When one keeps external referents in mind, it is meaningful to perform arithmetic on Dedekind cuts. And Dedekind was thorough in that regard, applying it to arithmetic operations, to ordering relationships, and to limiting processes. If one places Dedekind's work in a realitybased context, Dedekind has supplied a systematic way to work with a certain species of indirect *specifications* of irrational numbers, a way that extends, however trivially, to rational numbers, as well.

There is even a sense in which Dedekind cuts are numbers,

just as 110 is a number. They have, of course, one essential

difference: The figure 110 functions like a word; a Dedekind cut

functions like a description. Even so, both a Dedekind cut and a

decimal expression such as 110, *identify* specific quantitative relationships, ratios, in the world. Both the Dedekind cut and the

decimal expansion provide, arithmetically, systematic ways to

establish indirect measurement of quantitative relationships, by

exploiting the relationships between these quantitative

relationships. One of these ways, specifying a Dedekind cut, is

much more complex than the other, but, properly viewed and used,

both ways serve the same ends.

Cauchy sequences are a similar case. Earlier in this chapter,

I identified just how a Cauchy sequence *specifies* a quantitative relationship. Cantor and Heine did not invent Cauchy sequences

nor did they invent arithmetic operations on sequences. But they

did provide a systematic approach to these operations, just as

Dedekind did for Dedekind cuts. And when they defined an

equivalence relationship between Cauchy sequences, they were

isolating, precisely, the condition under which two Cauchy

sequences converge to the same number. A Cauchy sequence is not

just a number nor a number, as such. But when a Cauchy sequence

is used to *specify a particular quantitative relationship, to specify a particular ratio* between magnitudes, and when one defines

arithmetic operations on Cauchy sequence, then a Cauchy sequence

functions as a number, in the same way that a Dedekind cut can function as a number.
Finally, when mathematicians decided that either

Dedekind cuts or Cauchy sequences could be taken as numbers,

they were right, but, again, for the wrong reasons. In a realitybased, referential approach to mathematics, there are generally many ways to specify any particular quantitative or mathematical

relationship. And all of these ways are *specifications* of the *same* relationship. I discussed this principle earlier in my discussion of

√2. As in that example, it is *reality*; that ties them together. It is the fact that they all relate to the same existents, that they all identify

the same relationships, that ties together all the various

characterizations of the same external relationship.[54]
When mathematicians subsequently demonstrated a oneto-one relationship between Dedekind cuts and Cauchy sequences, they showed, in reality, in *my* context, how one kind of specification of ratio could be translated into another. For example, they showed

how the limit of a Cauchy sequence could be identified in the form

of a Dedekind cut, by specifying the relationship of the limit to the

rational numbers. In general, navigating between different ways of

specifying or identifying quantitative relationships is tremendously

important in mathematics. Every such navigation enhances one’s

ability to establish quantitative relationships, to perform abstract

measurement. It is one of the keys to indirect measurement, akin to

the importance of intersections in Euclid's *Elements*, mentioned in Chapter 1. In relating Dedekind cuts to Cauchy sequences, mathematicians confirmed that the two kinds of specifications were consistent. Take addition. From a referential, realitybased perspective, one adds Cauchy sequences to specify, in the form of a Cauchy sequence, the sum of the ratios that they represent. One adds two Dedekind cuts for the same reason: to specify, in the form of a Dedekind cut, the sum of the two ratios that these Dedekind cuts represent. Assuming that both measurements are correct, they *must* give the same answer because they *refer to the same external relationship*. If they failed to give the same answer, it would signal a mistake either in the formulation of Dedekind cuts or of Cauchy sequences. To show that the two formulations match is exactly what an accountant does when he calculates a balance in two different ways and compares the results. It's a confirmation, one way of detecting possible error.

The German mathematicians provided two ways to define real numbers and the mathematical community decided that it didn't matter which choice one made because they were formally equivalent. But, from a realitybased perspective, no choice is required for quite a different reason. First, there is positive value in both. Both provide systematic ways to identify ratios of magnitudes

both. Both provide systematic ways to identify ratios of magnitudes. Secondly, they do not compete with each other; they supplement each other. One does not, once and for all, choose one way of identifying a quantitative relationship over another; one leverages both. Finally, one relates the two formulations, not because form is all that matters, but because translation matters. And translation between two formulations is both possible and desirable when and because they both formulate the same underlying reality and provide different perspectives on that underlying reality.

Final Comments

Dedekind's work, and Cantor's use of Cauchy sequences, properly interpreted, exhibited the relationship of irrational numbers to rational numbers and gave meaning to the operations of arithmetic as they apply to irrational numbers. Restored, over Dedekind's dead body, to its proper referential context, Dedekind cuts, in particular provide a systematic way, a modern implementation of the fundamental insights of Eudoxus, to utilize rational numbers to characterize irrational numbers and to establish the domain of real numbers as comprising both rational numbers and irrational numbers.[55]

Cauchy's earlier work had established the definitions of

limits and continuity that were later formalized in the well-known

“epsilon-delta” approach to continuity and limits. Cauchy’s work

completes the theory of approximation by establishing the proper

definition of a limit: One says, for example, that a function $y = f(x)$

has a limit $y = y_0$ at $x = x_0$ if, no matter how close one needs to approximate y_0, one can guarantee the required precision by selecting an x sufficiently close to x_0.[56]These insights are essential and fundamental. But they are not sufficient to understand the way

that limiting processes relate to the world that they measure. So

they left a gap, a gap that Dedekind and Cantor rushed to fill later

in the century.

Properly interpreted, both Dedekind cuts and Cauchy

sequences provide valuable tools. As I have shown in my positive

treatment, Cauchy

sequences, in particular, provide infinite

mathematical precision by providing a system of approximations, a system that presupposes that all specific precision requirements are

finite, but provides a contingency for any specific precision

requirement that might ever arise.

At this point in the discussion, we still stand at the

threshold of mathematical abstraction. But as one proceeds to

study higher mathematics, the same issues resurface repeatedly.

For example to solve differential equations one needs to extend the

theory of approximation to include mathematical functions. Yet the essential principles and required understandings are encountered at the beginnings of the subject. The principles, though their application may require specialized knowledge, remain the same.

[1] Jeremy Gray, *Plato's Ghost the Modernist Transformation of Mathematics*, 2008, Princeton, Princeton University Press, Apparently Newton held this view: On page 134, Gray refers, in passing, to the "Newtonian view that the real numbers were ratios of quantities." See also: Penelope Maddy, *Realism in Mathematics* (Oxford, Clarendon Paperbacks, 1992), p 89, for an alternative viewpoint from a "set theoretical realist" perspective: "Knowledge of numbers is knowledge of sets, because numbers are properties of sets."

[2] As a young student, I recall the explanation that 2/7 + 3/7 = 5/7 because two of something plus three of something is five of something. For a discussion from a similar perspective, see Ronald Pisauturo and Glenn D. Marcus, *The Intellectual Activist*, September 1994, Vol 8, No 5, "The Foundation of Mathematics, Part II", p 10.

[3] Aristotle, *Physics*, edited by Jonathan Barnes, Revised Oxford Translation (Princeton University Press, 1984), Book III, Section 6, at 206^{b}, lines 12–13

[4] Ayn Rand, *Introduction to Objectivist Epistemology Expanded Second Edition*, April 1979, paperback edition, p 17–18

[5] Rand, p 201, "For instance, all you have as the basis of your operation is the arithmetic series. You don't need any further definitions as a base. From then on you work with that base."

[6] Pat Corvini, lecture entitled "Two, Three, Four, and All That," summer 2007, available on CD from the Ayn Rand Book Store (www.aynrandbookstore.com), offers a detailed reduction of integers and rational numbers to their perceptual roots. See also: Maddy, see also, p 179, "If there is nothing to decide between the von Neumann and the Zermelo ordinals when identifying the natural numbers with sets, how can either sequence of sets claim to actually be the numbers? The set theoretic realist's answer, implicit in the account of set perception, is that neither sequence is the numbers, that numbers are properties of sets which either sequence is equally well equipped to measure. The same line of response works for the real numbers when they are understood as detectors for the property of continuity."

[7] Archimedes, *The Works of Archimedes* (Cambridge: at the University Press, 1897), "The Sand Reckoner" p 221-232. Sir Thomas L. Heath, *A Manual of Greek Mathematics* (New York: Dover Publications, 2003),

pp 40–41

[8] Rand, p 196

[9] Pat Corvini, lecture entitled “Achilles, the Tortoise, and the Objectivity of Mathematics,” summer 2005, available on CD from the Ayn Rand Book Store

(www.aynrandbookstore.com), makes a closely related point about convergence

[10] Rand, “Definitions,” p 40, also, in same volume, Peikoff, Leonard, “The Analytic-Synthetic Dichotomy,” p 98, “*The meaning of a concept consists of the units – the existents – which it integrates, including all the characteristics of these units*.”

[11] Judith Grabiner, *The Origins of Cauchy’s Rigorous Calculus*, New York, Dover Publications, 1981, p 102-106

[12] G. W. Leibniz, *The Early Manuscripts of Leibniz*, translated by J. M. Child (New York: Dover Publications, 2005), “History and Origins of the Differential

Calculus,” p 23

[13] Grabiner, p 102–106, footnote 83 on p 208

[14] Joseph Dauben, *Georg Cantor His Mathematics and Philosophy of the Infinite*, Princeton University Press, 1990, p 165-168

[15] Dauben, p 120-132. also, *A History of Analysis*, edited by Hans Niels Jahnke (Rhode Island, American Mathematical Society, 2003 hardback), Moritz Epple,

Chapter 10 “The End of the Science of Quantity: Foundations of Analysis, 1860 –

1910.” p 120-132

[16] Grabiner, p 102–106, footnote 83 on p 208

[17] Carl B. Boyer, *The History of the Calculus and Its Conceptual Development*, Dover Publications, Inc. Mineola, New York, 1949, p 281, Boyer points out the issue, as follows: “... one cannot define the number √2 as the limit of … because to prove that this sequence has a limit one must assume … the existence of this number …”

[18] Corvini, “Achilles,” makes a very similar point about the limit of a sequence

[19] Epple, p 300

[20] Technically, this approach would need to be modified to address terminating

decimals such as 1.2. The issue is that 1.2000 … = 1.19999 … . So the decimal 1.2

might be given by the sequence 1.19, 1.201, 1.199, 1.2001, 1.999, 1.20001, …

[21] See note 221

[22] Pat Corvini, lecture entitled “Two, Three, Four, and All That: The Sequel,”

summer 2008, available on CD from the Ayn Rand Book Store

(www.aynrandbookstore.com), offers an account of the real numbers from a similar, realitybased, perspective

[23] Rand, p 305

[24] Rand, p 35

[25] Rand, p 36

[26] Carl B. Boyer, *History of Analytic Geometry*, Dover Publications, Inc.

Mineola, New York, 1956, Chapter V, p 74–102

[27] Euclid, *Elements*, Book V, Archimedes. *The Works of Archimedes*, "On Floating Bodies," p 253–300

[28] Compare Corvini, "Achilles"

[29] Corvini, "Achilles"

[30] Corvini, "Achilles"

[31] Boyer, *History of Calculus*, p 281

[32] Gray, p 16

[33] Gray, Chapter 3, "Modernism Arrives", p 113-175. Epple, Chapter 10 "The End of the Science of Quantity: Foundations of Analysis, 1860 – 1910"

[34] Gray, section on "What Are Real Numbers?" p 129-135 and section on "Peano", p 168-170

[35] Richard Dedekind, *Essays on the Theory of Numbers*, Dover Publications 1963 from a 1901 English translation, German publication 1872, p 9-10

[36] Maddy, p 84-86, provides a telling critique of this ontological approach

[37] Dedekind, p 9-10

[38] Dedekind, p 13

[39] Morris Kline, *Mathematical Thought From Ancient to Modern Times*, Volume 3, Oxford University Press, Oxford, 1972, 1990 paperback edition, p 986, gives

Dedekind's own view on this point: "[Dedekind] should say that the irrational

number *a* is no more than the cut. In fact Heinrich Weber told Dedekind this, and in a letter of 1888 Dedekind replied that the irrational number a is not the cut itself but is something distinct, which corresponds to the cut and which brings about the cut."

In fact, a Dedekind cut is a distinction, distinguishing a particular number. But a distinction needs an object to be meaningful. Dedekind does not supply an object. So the only non-Platonist answer in sight is the means of making that distinction,

namely the two sets by which the distinction is made.

[40] Dedekind p 21

[41] Dedekind p 15-17

[42] Dedekind p 22

[43] Stewart Shapiro, *Thinking about mathematics* (Oxford, Oxford University Press, 2000 paperback), p 258, "The essence of a natural number is its relations to other natural numbers." This appears to be a distinguishing tenet of structuralism.

[44] Dedekind p 39

[45] Dedekind p 39-40

[46] Dedekind p 40

[47] Dedekind p 9-10

[48] Dedekind p 10

[49] Corvini, “Two, Three, Four, and All That: The Sequel,” offers a somewhat different critique from a similar general perspective. Also Maddy, p 81-86 elaborates a point by Benacerraf to argue that real numbers are not sets. She concludes, “Of course, if we take Benacerraf’s argument that natural numbers aren’t sets to be

persuasive, as I think we should, an analogous line of thought shows that the reals can’t be sets either. We could tell a story of Georgie (for Georg Cantor) and Rich (for Richard Dedekind), one of whom learns that the reals are Dedekind cuts and the

other of whom that they are Cantor’s fundamental sequences. The rest of the story follows as before, and the conclusion: real numbers aren’t sets.” Kline, p 986,

quoting Dedekind “[the] irrational number *a* is not the cut itself but is something distinct, which corresponds to the cut and which brings about the cut.” is also

relevant here

[50] Dedekind p 3

[51] Gray, p 131-133. Epple, p 295-301

[52] Epple, p 300

[53] Stewart Shapiro, p 258, again, “The essence of a natural number is its relations to other natural numbers.”

[54] Maddy, is on a very similar track and even brings in the issue of measurement.

Regarding natural numbers, she writes, p 89, “The choice between the von Neumann

and the Zermelo ordinals is no more than the choice between two different rulers

that both measure in metres. The debate between Ernie and Johnny is like an

argument over whether an inch is wooden or metal.” Even more to the point, on p

94, “… what makes one set theoretic version of the reals preferable to the others?

Answer: nothing; each version serves to detect and measure the same underlying

properties.”

[55] Dedekind

[56] Grabiner, pp 93-112

The following pattern was known to Pythagoras:

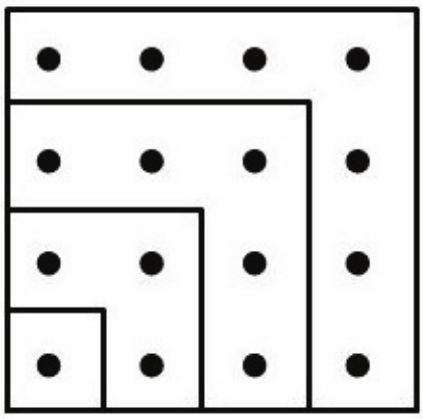
Chapter 5 Geometry and Human Cognition

Introduction

The following pattern was known to Pythagoras:[1]

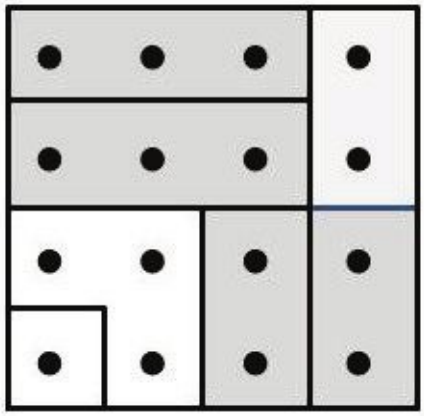
$$1 = 1 = 1 \times 1 = 1^2$$
$$1 + 3 = 4 = 2 \times 2 = 2^2$$
$$1 + 3 + 5 = 9 = 3 \times 3 = 3^2$$
$$1 + 3 + 5 + 7 = 16 = 4 \times 4 = 4^2$$

It turns out that this pattern is no coincidence, is true in general, and can be proven as a general fact by algebraic methods. But this is neither the way it was discovered nor validated by the Greeks. In essence, they saw it directly from the following picture:



Think of the figure as being built in stages from the bottom left corner. Start with the smallest square consisting of one dot. Then add an L-shaped figure (known as a “gnomon”) of three dots to form a two by two square. Next add five dots to form a three by three square. And so on. Each gnomon has two more dots than the previous one and one is left each time with a larger square. Since the process starts with one dot, the series of dot counts in successive gnomons is the series of odd numbers. Once you’ve grasped the process, the entire generalization is captured in one picture.

If it’s not immediately obvious that successive gnomons differ by two dots, a further elaboration should help. Namely, matching dots for two successive gnomons, yields:

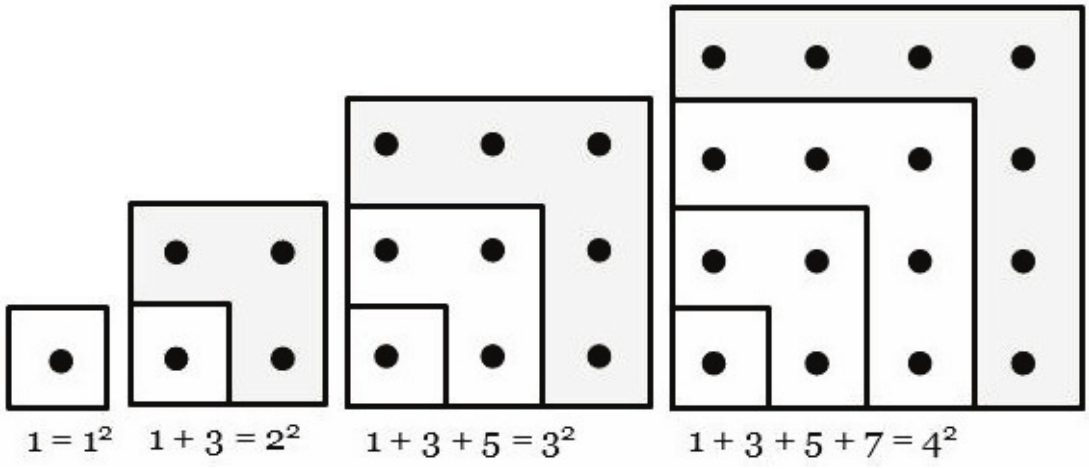


The top dark grey box matches three dots of the outer gnomon to three dots of the previous one. The dark grey box on the right matches two dots of the outer gnomon to the remaining two dots in the previous gnomon. When this is done, there are two remaining dots in the outer gnomon. These dots, located in the light grey boxes, are left unmatched.

On reflection one sees that the same pattern would hold no matter how many gnomons one might add, without regard to the size of the square.

What can one learn from this? First notice the way that the first figure gave perceptual reality to an abstract relationship. One looks at the figure; one grasps the relationship.

Now this does not happen automatically. One needs to regard the picture as being built in stages. It would have been helpful, for example, to see a sequence of pictures:



In effect, if one is able to grasp the relationship from the very first diagram, it is by going through this sort of sequence in one's head. One performs a process to grasp mentally the abstract relationship illustrated or embodied in the diagram. The diagram is as stylized as possible to point one in the right direction.

The picture is just a picture, but it is a picture of an abstract relationship. It is a concretization of that relationship. With the right mental focus, it enables one to see, in one concrete, a generalization that holds no matter how far one goes in the series, no matter how many gnomons have been added. And once the integration is complete, one sees it as an embodiment of a universal principle.

One does not require a formal algebraic proof to grasp the relationship as a universal pattern. This stylized geometric representation conveys an understanding of a mathematical principle that no algebraic derivation could equal.

I've now mentioned stylization twice. How have I stylized this picture? First, I made all of the dots the same size, color, and shape, because my intent is to regard them as units of a particular kind. Differences in size, color, or shape would have created a distraction from the essential relationships. I spaced them evenly for the same reason. I made an effort to make the vertical spacing the same as the horizontal spacing. I included nothing in the picture that was not relevant to the point it was supposed to convey.

The effect of such a stylization is to treat attributes such as size, color, shape, and spacing as omitted measurements, as irrelevant to the pattern, as potential distractions to, therefore, be ignored. In the end, the figure is an organized whole that, in totality, conveys the intended abstraction.

The use of stylization to visually convey an abstraction is not unique to this example, nor unique to mathematics. It is one of the ways we present and concretize abstract ideas. An organization chart, for example, captures a complex set of human relationships on one sheet of paper, providing a way to quickly zero in on any particular detail without losing sight of the whole.

The power of stylized representations to crystallize a conception is important generally, but that power, and its personal importance, is most striking in the visual arts. And stylization, a focus on the essential, is what makes it possible. As Ayn Rand explains:
"…an artist isolates the things which he regards as metaphysically essential and integrates them into a single new concrete that represents an embodied abstraction."[2]

and

"The so-called visual arts (painting, sculpture, architecture) produce concrete, perceptually available entities and make them convey an abstract, conceptual meaning."[3]

Finally, discussing a stylized painting of some apples, she concludes with the stronger statement:
"What is it, then that the artist has done? He has created a visual abstraction."[4]
A longer excerpt will help elucidate her context:

"It is a common experience to observe that a particular painting – for example a still life of apples – makes its subject "more real than it is in reality." The apples seem brighter and firmer, they seem to possess an almost self-assertive character, a kind of heightened reality which neither their real-life models nor any color photograph can match. Yet if one examines them closely, one sees that no real-life apple ever looked like that. What is it, then, that the artist has done? He has created a visual abstraction.

He has performed the process of conceptformation – of isolating and integrating – but in exclusively visual terms. He has isolated the essential, distinguishing characteristics of apples, *and integrated them into a single visual unit*. He has brought the conceptual method of functioning to the operations of a single sense organ, the organ of sight."[5] (Emphasis, mine)

One can distinguish two broad parallels between this square-of-dots example and

One can distinguish two broad parallels between this square-of-dots example and Ayn Rand's theory of the cognitive role of art. First, in what is achieved: increasing ones grasp of abstract relationships by integrating them into concretes. And second, in how it is achieved: through a process of stylization, directing one's attention to certain essentials of the abstraction it is designed to convey. The geometric illustration is, on the one hand, a concrete embodiment of the general principle. But it conveys this wider principle insofar as it enables the viewer to isolate the wider principle, to perceive the operation of that principle within the particular concrete, and to realize that the principle in no way depends upon the specific dimensions of the particular square nor the particular arrangement of the dots, nor the particular object being counted that the dots might represent.

Regarding stylization, my central interest in this example is what it illustrates about the role of geometry in mathematics. Stylization in the square-of-dots example consists in simultaneously *capturing* and *isolating* the abstract relationship it is intended to convey. And this aspect is common to all applications of geometry to embody mathematical relationships.

Throughout this book, I have emphasized two complementary and inextricably bound perspectives on mathematical phenomena. Measurement, the essential purpose of mathematics, ultimately involves an *identification* of a quantitative relationship between *concretes* in the world. The very *process* of measurement involves a relationship, a relationship between the identification, including the *means* of that identification, versus the *objects* one is relating quantitatively. This is a relationship between the means of measurement and the objects of measurement, objects viewed from an abstract perspective.

When I discussed numbers, as in Chapter 4, I focused on a system of measurements by which specific relationships can be identified, but independently from any particular concrete. And here I appeal, once again, to Ayn Rand's theory of omitted measurements: the meaning of the number does not depend on any one specific concrete to which it might apply, but pertains equally to all of these concretes. This is the focus on the *means* of measurement.

When I provided an account of magnitudes and on the relationships between them, as in Chapter 2, I focused on the *objects* of measurement. I did not treat them as objects *apart from* measurement, but, rather, as the objects *of* measurement, as objects relatable quantitatively. I viewed them from an abstract perspective, as representative of a broad category of existents to which one can

apply a specific measurement, either a specific relationship to a standard or a specific relationship between two concretes and embodied in those concretes. This is the geometric perspective.

Identification versus object identified: This is the contrast between Chapters 2 and 4. Chapter 2 presented a geometric perspective on magnitude, the object of measurement, and Chapter 4 treated numbers, as comprising a system of measurements to identify relationships between magnitudes. This latter is the focus on the means of measurement or, more precisely, on one key aspect of that means: not the physical process, but the system of concepts by which a specific relationship is named, by which it becomes a mental unit that can be retained and distinguished from other similar mental units naming different quantitative relationships.

Geometric objects, in general, are *actual* objects and relationships in the world, viewed from an abstract perspective. By means of geometry, one isolates the essential features of certain relationships in the world while retaining a perspective, as well, on the *objects* that exemplify those relationships.

Geometry is a concretization of quantitative relationships that, in general, transcend *spatial* relationships. Yet, at the same time, geometric objects are abstractions that integrate and embody a vast range of concretes in the world. When one focuses on the concrete meaning of a mathematical relationship, one's perspective is geometric.

The square-of-dots example embodies two aspects of geometry that go beyond these general observations. First, geometry is not just about shapes and spatial relationships. Second, one can utilize geometrics means, including visualization, to exemplify general mathematical relationships between numbers and, more broadly, between concrete measurements of any sort within a system of measurements. I add this last elaboration because numbers are only the *most* important, not the only important system of measurements in mathematics. For example, Chapter 8, introducing group representations, treats an important category of measurement unrelated to magnitudes, as such.

The square-of-dots example is a concretization of a mathematical relationship. The stylization in this example is a way of isolating and focusing one's attention on the aspects of that concretization that pertain to the relationship one intends to exemplify, a concretization that makes it possible to grasp a broad principle by a

focused attention to a single example. And, although my focus in this short chapter, expanding a major theme throughout the book, is on the role of geometry in mathematics, the square-of-dots example illustrates, as well, the cognitive role of a well-chosen *example* in mathematics. But, to return to my theme, one of the things that this particular example illustrates is that geometric examples can be applied, in general, to mathematical relationships. And the deepest reason for this is that all mathematical relationships pertain ultimately to relationships among concretes.

The essential role of geometry in mathematics is to conceive of a constellation of mathematical relationships as embodied in an object, an object taken to exemplify, yet embrace, the wider class to which the constellation applies. The cognitive role of the geometric perspective, its ability to concretize an abstraction, is one key to the unique, powerful, and fascinating role that geometry, throughout the history of mathematics, has played in mathematics and in the fields that use mathematics.

I spent two chapters, Chapters 1 and 3, expounding my perspective on Euclid's system of geometry, as it pertains to plane figures. Chapter 2, following Euclid's general approach, provided a geometric treatment of magnitude. The purpose of the current chapter is to further delineate the role that geometry plays in mathematics. To that end, I challenge the naïve views that geometry is just about two and three dimensional objects and that its value consists exclusively in providing pictures to support geometric arguments. Geometry does indeed study shapes; pictures are indeed important. However, these do not exhaust the domain of geometry nor its role in mathematics, nor its role in the applications of mathematics.

Beyond Plane and Solid Geometry Magnitudes

Geometry is not limited to the study of plane and solid figures and it was not so limited even in Euclid's *Elements*. Numerous chapters (called Books) of Euclid are devoted to studying relationships between magnitudes or between numbers. Euclid's treatment applies purely geometric methods to both magnitudes and numbers, representing both magnitudes and numbers by line segments.[6] Such segments are usually depicted horizontally, occasionally vertically, but it is clear from the outset that the orientation of these segments is irrelevant to the discussion and, from our vantage point, to be considered an omitted

measurement.

The magnitudes represented by these line segments in Euclid might be thought of as lengths, areas of geometric shapes, volumes of geometric solids, or whole numbers, though numbers, Euclid's viewpoint notwithstanding, are not magnitudes. But the magnitudes found in Euclid hardly exhaust the possibilities. In a physical context, weight, mass, velocity and acceleration are all magnitudes. Already in Greek antiquity, one finds Archimedes using line segments to represent force in his study "On Floating Bodies".[7]

The most important goal of Euclid's studies of both numbers and magnitudes is to develop a theory of ratio. The theory of ratio plays a major role in Euclid and one finds, in Euclid's *Elements*, ratios of lengths, of areas, and of numbers. One also finds Euclid expressing equalities between a ratio of lengths versus a ratio of areas and versus a ratio of numbers. In regard to magnitude, following Eudoxus, Euclid goes to considerable length to justify such equalities.[8]

But one never finds Euclid taking a ratio of a length to an area. And this is for a very good reason. Ratios, in Euclid, presuppose an ability to compare the sizes of the two participants in the ratio. But, one cannot, with respect to size, compare a length to an area.

To make this concrete, consider a square that is one foot long in both directions. Is the length of the side greater or less than the area? Suppose, for example, that one attempts a numerical answer. Well, if one's unit is inches, the length of the side is 12 inches and the area is 144 square inches. One might observe that 144>12and betempted toconcludethat theareaisbigger,butone would still be comparing apples and oranges. For the numerical answers and their rankings depend totally on one's choice of units. For example, notice that, measured in yards, the length of each side is 1/3 and the area of the square, in square yards, is 1/9. As numbers, 1/9 is, clearly, less than 1/3. In sum, by one measure the area is greater; by the other, it is smaller. One cannot, then, compare the size of an area to the length of a line.

Although one cannot add and subtract magnitudes of different kinds, it is commonplace *today* to divide miles by hours to get miles per hour. But this ability, taken for granted today, was dearly won and involves choices of units, such as miles and hours, for each factor. Such ratios cannot be found in Euclid's *Elements*. Euclid's geometric study eschewed all consideration of units and his

definition of ratio was strictly and explicitly limited, was only applicable, to ratios between magnitudes of the same kind. Euclid could define ratios between lengths and ratios between areas. He could even equate a ratio between areas to a ratio between lengths. But he simply could not conceive of a ratio of a length to an area. For more on this point, recall my discussion in Chapter 2.

In general, Euclid treats magnitudes as he treats triangles. He treats them as abstract entities that can be related to each other in various ways without explicit reference to a chosen unit. For the relationships he uncovers do not depend on such choices. In this, Euclid takes an abstract perspective on physical magnitudes whose mathematical relationships to each other exist independently from any particular measurement or choice of unit. The line segments that Euclid uses to represent magnitudes are visual abstractions, used to provide a concrete reference point to support his geometric arguments.

Cartesian Coordinates

If you have ever seen a graph, you have seen the use of Cartesian coordinates. Graphs are everywhere in today's world. For example, historical movements of stock prices, revenue, and profitability are all routinely displayed graphically. There is a horizontal axis, usually considered the x axis (measuring time in these examples). And there is a vertical axis (usually considered the y axis) representing the particular quantity being measured and compared across time or across whatever variable the x axis represents. Corresponding to each axis is a unit representing the quantity measured by that axis. The specifications of the x and y axis, the zero point of each axis, and of the units in both directions, all collectively constitute the set of Cartesian coordinates.

Descartes introduced Cartesian coordinates as part of his program to reduce geometric questions to algebraic questions.[9] In his treatise, Descartes derived formulas for straight lines, circles, and other geometric figures. Much of his work consisted in relating his approach to the study of conic sections and other shapes by the Greeks.[10]

The most important relationship between the Greek and Cartesian perspectives is the Cartesian use of algebraic equations to find the intersections of geometric shapes. When one uses coordinates, one finds the intersection of a line and a circle by finding the solution of two equations in two unknowns. One begins

with the two equations that define, respectively, the circle and the straight line in question. The unknowns in each of these equations are the x coordinate and the y coordinate.

For example, $3y = 4x$ is the equation of a line and $x^2 + y^2 = 25$ is the equation of a circle. A pair consisting of x and y coordinates falls on the line precisely when their substitution into the equation of the line produces an equality. Thus, the point (x, y) = (9, 12) or, equivalently, x = 9 and y = 12, is a point on the line because 3 times 12 equals 4 times 9. Similarly, the point (5, 0) is on the circle. In such cases the coordinate pairs are said to satisfy the equation of the line or, respectively, the circle. A pair of x and y values can be an intersection point only if it simultaneously satisfies both equations. In this example, the points (x, y) = (3, 4) and (x, y) = (-3, -4) satisfy both equations and identify the two points at which the line and the circle intersect.

Prior to Descartes, the Western world used Euclid's *Elements* as its mathematical framework to study question in physics. This can be seen in *Two New Sciences*[11] by Galileo and even Newton's *Principia*,[12] written after the time of Descartes. But ultimately, the tables were turned (so to speak) and geometric questions began to be formulated and studied algebraically.

However, the relationship works in both directions. Through these very same coordinates, any problem in algebra can be regarded as a problem in geometry. For example, the equation $x^2 - 3x + 2 = 0$ can be thought of the values of x for which the graph of $x^2 - 3x + 2 = y$ intersects the x axis (characterized by y = 0). These two points have coordinates (2, 0) and (1, 0), corresponding to the two solutions x = 2 and x = 1 to $x^2 - 3x + 2 = 0$.

Whenever the focus is on the kind of *object* (or constellation of quantitative relationships as embodied in an object) being measured, one thinks of it geometrically. One takes an abstract perspective on concrete embodiments of the mathematical relationships, one treats the geometric object as an entity, as a particular unit treated hypothetically, as an instance of a universal constellation of relationships.

On the other hand, whenever the focus is on the *measurements* of the object, one uses Cartesian coordinates. In the first case the focus is on the object as existing independently of its measurements. In the second case, the focus is on the

algebraic and functional relationships among the coordinates without specific regard for an external object. In the centuries since the introduction of Cartesian coordinates, the integration and interplay of these two perspectives has become a fundamental pivot point in mathematics and is implicit in any application of the mathematics.

I have already illustrated that geometric figures do not always represent geometric shapes. The Cartesian framework carries this one step further. Cartesian coordinates may be used to represent the Euclidean plane, but they need not. In the stock market example, neither the x dimension nor the y dimension is a spatial dimension. Moreover, the kind of quantity represented by the y axis, in this and in countless other examples, is different than the kind of quantity represented by the x axis.

In such cases, the functional relationship depicted by the graph is not a shape in the strict Euclidean sense, but it should still be viewed as an abstract relationship that exists independently of the particular coordinates used to represent it. For example, if one changes the zero point (the starting reference date) of the x axis or measure revenue in million-dollar increments instead of thousanddollar increments, the relationship being expressed hasn't changed. Only the expression of that relationship has changed.

You may remember polar coordinates from high school. Polar coordinates represent a point by a pair of coordinates traditionally denoted by 'r' and 'θ', representing, respectively, the distance from the origin and the angle its line from the origin makes with the x axis. Descartes did not have polar coordinates. But their use provides one more example of my last point. The circle that is expressed in Cartesian coordinate as $x^2 + y^2 = 25$ is expressed in polar coordinates as $r = 5$. In making this comparison one conceives the circle as existing independently of the means used to express its shape.

As a further ramification, there is no essential limit to the number of Cartesian coordinates one might have reason to employ. For example, consider the motion of a ball through the air. One assigns a spatial position of the ball, which takes three coordinates, to each instant in time – which requires a fourth coordinate. The order here is optional. Instead, of taking it last, one could logically take time as the first coordinate since the three spatial coordinates should be regarded as depending on time.

Even more coordinates would be required to describe the motion of a pencil

Even more coordinates would be required to describe the motion of a pencil thrown into the air because one would want to capture not only the ever-changing position of the pencil, but also its orientation as it hurtles through space. These coordinates can be thought of as expressing a position in the "configuration space" for the pencil, meaning, by "configuration space" the universe of potential positions and orientations of the pencil. Looked at geometrically, the motion of the pencil is a path in its configuration space. Looking at it geometrically means that one focuses on its path as a phenomenon in the world, existing independently of the choice of coordinates used to measure the path, just as Euclid spoke of magnitudes without specifying a particular unit. This perspective, the unity of the object, is put into the background when one simply expresses its path as five or six independent functions of time. (Six if you include the spin of the pencil around its long axis.)

As it relates to the world, the configuration space takes an abstract perspective on the concretes that are the meaning and referents of the concept. But, in relation to one's grasp of the complex of relationships, the configuration space serves as a concretization of an abstract relationship, a concrete that is taken as symbolizing and exemplifying an abstraction, in this case the possible positions and orientations of a three dimensional object. In this, it functions in a manner that parallels the role that Ayn Rand finds for visual abstractions in art.

Clearly, I've now gone beyond the limits of a visual representation of a geometric shape. Yet the geometric conception still applies whenever one thinks of a trajectory abstractly as having an existence independent of the particular coordinate axes used to measure the motion. Whenever physicists speak of changing coordinate systems, as they must, they are presupposing and distinguishing the independent existence of the phenomena they capture in their abstract formulations from the means that they use to capture it.

With these increasing dimensions, one leaves visual abstractions behind, but one does not thereby leave behind the more fundamental conception of an abstract entity, of a geometric abstraction. But now one expresses these geometric abstractions in purely conceptual terms, just as a novelist uses concepts to depict the characters in a novel. The artistic analogy still applies, but now an analogy to literature must replace the analogy to the visual arts.

The cognitive role of geometry is to conceive of a constellation of mathematical relationships as a unity, as embodied in an abstract geometric object, one that provides an abstract perspective on and refers to concretes in the world. One

begins to do this with visual abstractions and one derives enormous benefit from the ability to visualize complex relationships. Even when the number of variables exceeds 3, there is still a kind of visualization that can sometimes aid one's grasp of these relationships. However, the deeper need for geometry, creating a single unit to embody an organized constellation of mathematical relationships, transcends one's ability to draw pictures and even to visualize.

Final Remarks

The applicability of these ideas is everywhere in mathematics because the geometric perspective is always in the background even when it is not explicit. These ideas apply to, and provide a perspective for, the entire history of mathematics. Geometry serves a cognitive need and, at the same time, geometric abstractions provide the bridge between concepts of measurement and their ultimate referents in the world. In this outline of those relationships I have continued a thread that runs through the entire course of this book.

[1] Sir Thomas Heath, *A History of Greek Mathematics, Volume I From Thales to Euclid*, Dover Publications 1981, p 77 in Chapter III "Pythagorean Arithmetic"

[2] Ayn Rand, *The Romantic Manifesto, Revised Edition*. Signet, 1975 (paperback edition), Chapter 1, "The Psycho-Epistemology of Art", p 20

[3] Rand, p 47 from "Art and Cognition"

[4] Rand, p 47-48

[5] Rand, p 47-48

[6] Euclid, Books V and X concern magnitudes, Books VII – IX concern numbers

[7] Archimedes, *The Works of Archimedes*, 1897, Cambridge: at the University Press, "On Floating Bodies," pp 253-300

[8] Euclid, Book V

[9] Rene Descartes, *Des matiers de la Geometrie*, 1637, available in English translation as *The Geometry of Rene Descartes*, Dover Publications, Inc., 1954

[10] Especially, Apollonius whose book on conic sections is available in English translation as Apollonius of Perga, *Conics*, Green Lion Press; new rev. ed edition (October 1, 1998)

[11] Galilleo Galilei, *Discorsi E Dimonstrazioni Matematiche intorno a due nuoue Scienze*, 1638, English translation as *Dialogues Concerning Two New Sciences*, Northwestern University, 1946

[12] Sir Isaac Newton, *Philosophia Naturalis Principia Mathematica*, 1686, available in English translation by University of California Press, 1962

PART 2: ADVANCED

Chapter 6

Set Theory and Hierarchy in Mathematics

Introduction

Whatever its merits as a foundation for mathematics, the language of set theory is common currency in mathematics. The language of set theory offers and provides a generic set of concepts and notation that apply to any system of measurements or to geometric structures of any sort. Mathematical sets are a way of isolating, of zeroing in on something: solutions to an equation, a range of numbers, or a figure in the plane.

But the waters get deeper whenever new domains of mathematical study are identified. A new mathematical domain, a new system of measurements or some new kind of geometric structure of any sort, regardless of how it actually arises in mathematics, is typically characterized as “a set for which …”, short circuiting any discussion of how the new mathematical domain arises in the first place. Such discussions may indeed be provided,

as motivation and good pedagogy. But, aside from the mandatory inclusion of examples, definitions in terms of sets are generally taken to stand on their own, independent of such motivation.

Set theory has its roots in the last decades of the nineteenth century, in the work of late nineteenth century mathematicians such as Cantor and Dedekind.[1] By the mid-twentieth century, the use and abuse of set theory was taken for granted.

Its abuse was epitomized by a group of mathematicians writing under the pseudonym of Nicolas Bourbaki.[2] Taking sets as their starting point for every branch of mathematics, they remained deliberately silent on what these sets might be sets of.

But this silence among mathematicians was general. In a classic text intended to provide a basic working knowledge of set theory, a prerequisite for studying advanced mathematics, Paul Halmos begins his second paragraph with "One thing that this development will not include is a definition of sets."[3] Paul Halmos was not a Bourbaki and was noted for his ability to clearly motivate and explain concepts in advanced mathematics.

In the Bourbaki approach, the introduction of any specialized study would typically begin with a definition involving a set of a specified type. As a typical example, J. Dieudonne, a prominent Bourbaki, in his classic *Foundations of Modern Analysis*, having spent most of a page giving an unmotivated,

formal definition of a *distance* (a generalization of the colloquial term), provides, without motivation, the definition, “a *metric space* is a set E together with a given distance on E.”[4] Dieudonne follows this definition with two pages of examples, showing that these

examples do, indeed, satisfy his definition.

The issue, here, is not whether the generalized concept of

distance is either valid or important. Indeed, to me, neither its

validity nor importance is in question. What’s distinctive is the

approach to exposition and the implications of such an approach.

In its practice of presenting examples only after giving a formal

definition of this type, the Bourbaki approach suggested that

motivating its concepts, was not particularly important, not a

requirement of definition.

In effect, they presented definitions without a genus, with

the undefined set taking the place of a genus, of an actual universe

of discourse. Reasoning from this base consisted of appeal to the

socalled axioms of set theory and, beyond that, to whatever special

properties, of any particular kind of set, that were asserted in the

definition. Calling something a set did indeed provide a warrant to

reason in a certain way without immediate fear of contradiction,

but it said nothing about what any of it might mean.

In this fashion the Bourbaki’s, emulating Hilbert’s famous

axiomatization of Euclidean geometry,[5] aimed to systematize and axiomatize mathematics. Bourbaki's influence began to wane

around 1970,[6] but much of the legacy they did not invent, but helped perpetuate, notably the foundational role of set theory in

mathematics, remains.

Within a reality-based account of mathematics of the sort

that I am outlining in this book, to characterize a mathematical

domain as a kind of set admitting a prescribed structure is exactly

backwards. A

set of mathematical objects, if such

exists,

presupposes a *mathematical domain* from which the mathematical objects are taken. If the language of set theory is valid within

mathematics at all, it must presuppose, in every case, a particular

universe of some sort of mathematical objects, relating to the

world, to which the language of set theory can be applied.

Mathematics, like any other science, requires a hierarchical

development with reality as its starting point and its motivation.

Unfortunately, as I elaborate later in this chapter, set

theory, as presented today, as a formal mathematics study, is a

floating abstraction with arbitrary, if carefully chosen, axioms and

strained constructions.[7]

Should set theory, then, be dismissed out of hand? Or is there a need for such a concept in mathematics? If so, what function does it fulfill and what context does it presuppose? Is it possible, and is it reasonable, to rehabilitate the mathematical concept of *set*, as a specifically mathematical concept?

What is a Set?

The concept of man applies to all men that have ever existed, exist today, or will exist in the future. The concept of man is openended.[8] But the referents of the concept man do not constitute a set. Nor is the concept of man at all unusual in that regard.

One can, however, talk about a set of dishes or the set of books that one owns. In general, the word “set” meaningfully applies to existents of some kind that have been specifically circumscribed and delimited. To group existents into a set is to distinguish them, based on the recognition that, for some purpose, however transitory it may be, the existents belong together in some way. As the simplest example, when one counts something, one ascribes significance to the objects that one is counting, taken as a whole. Implicitly, one views the collection as a set.

In such usage, elements of sets are treated as concretes,

concrete instances of the kind of thing that they are. A set of numbers of some kind consists, first of all, of specific numbers. There must be no ambiguity concerning either the kind of existents that are included or the status of any particular existent of that kind as belonging or not belonging to the set. To isolate a set is, first of all, to distinguish and isolate its members.

As applied to concretes in the world, there are no infinite sets. But there are, it is claimed, infinite sets in mathematics. Is there something special about mathematics?

Let's review some mathematical examples.

Consider the polynomial equation

$x^3 - 6x^2 + 3x + 10 = 0$

This third order (cubic) polynomial turns out to be factorable: One has

$x^3 - 6x^2 + 3x + 10 = (x + 1)(x - 2)(x - 5) = 0$

So its solution set consists of three numbers, namely, -1, 2, and 5. Looking at the solutions as constituting a set focuses on these solutions as isolated instances of a totality.

There are many contexts in which the

nature of the

solution set is more important than the actual solutions in any

particular case. For example, a cubic equation with real coefficients,

such as this, can never have more than 3 solutions and always has

at least one real number in its solution set. And this simple

observation is not an isolated curiosity, but a first introduction to a

complex study known as algebraic

geometry, a branch of

mathematics that grew out of the study of polynomial equations.

A somewhat more complex function is the sine function

that arises in trigonometry and assigns a number, ranging from -1

to 1 to every angle. If one conceives of an angle as measuring a

rotation, then, as one continues to rotate a line segment attached to a point, one returns to one's starting point and traces the same

directions over and over again. At 360^0, a rotating line returns, for the first time, to its original position and the value of the sine

function returns to the value it has for an angle of zero degrees.

Mathematicians do not normally measure angles in

degrees. Rather, they utilize a particular magnitude that is related

to the degree to which a line has been rotated. Specifically, a line

segment of unit length traces out, as it rotates, a circle of unit

radius. The length s of the arc (the *arc length*) traced out on the circle measures

the amount of the rotation. When one measures

angles in this way, one says that the angle is measured in *radians*.

Since the circumference of the unit circleis2π,arotation of $360^0$is equal to 2π radians. In these units π is equal to 180^0.

The function sin(s) = 0 is satisfied precisely when s = n

π

for integer n, precisely when the rotating line, rotating about the

origin (0, 0) in a counter-clockwise direction with its endpoint

starting at (x,y) = (1,0), crosses the x axis. So the set of solutions is

infinite; there is a solution, namely nπ, corresponding to each

integer n. When one says that the solution set is {nπ, where n is an

integer}, one simply expresses the fact that any value of s of the

form nπ (n an integer) will satisfy the equation. To consider these

solutions as belonging to a set is simply to adopt a different

perspective on the same fact, a perspective that focuses on the

solutions as a totality and as distinguished among the wider

domain of the real numbers, from real numbers that do not satisfy

the equation.

What is most significant about this solution set? First,

successive solutions, in both directions, are evenly spaced and,

second, more importantly, there is no limit to its potential extent.

Before I continue, notice that every member of this set is an

Before I continue, notice that every member of this set is an

irrational number. In chapter 4, I argued that irrational numbers

are *only* meaningful, meaningful as *irrational* numbers, in an abstract setting, in which they apply to an entire openended

category of measurements. I argued, further, that irrational

numbers are *concepts of method*. They are the way that one

preemptively tracks any quantitative distinction among magnitudes

that might, someday, in some context, be required in regards to

some concrete. If sets of real numbers are meaningful then *set*, too, must be a concept of method: A set maintains its distinctions on an

abstract level, in the same way as the number system does and for

the same reason, distinctions that go beyond the requirements and

capability of any specific concrete case, but are needed to apply

abstract measurements universally. It is possible, as a methodical

procedure, for a *set* of real numbers to make such distinctions because these distinctions are already embodied in the real number

system.

As a third example, consider an interval between 15.9 and

16.1. One might encounter such an interval in the course of a

measurement. Suppose the resulting measurement is 16, plus or

minus .1. One would say that the length, say, of something is

between 15.9 and 16.1 centimeters.

A mathematician would not hesitate to characterize the

potential values within that interval as a set of numbers. Any number that one could name is unambiguously either contained in the interval, is between 15.9 and 16.1, or it isn't.

Sets of this type, consisting of small intervals, are, as in this example, very often invoked as a way of specifying a degree of approximation.

Now consider the graph of the equation $x^2 + y^2 = 25$. One is interested in pairs of numbers that satisfy the equation, pairs such as $x = -3$ and $y = 4$, alternately expressed as $(x,y) = (-3,4)$.

Considered mathematically, there are a mathematically infinite number of solutions to this equation.

A mathematician would characterize the points on this graph as a set. Without ambiguity, a pair of real numbers is either a solution to the equation or it is not. A point in the plane, as represented by a pair of coordinates, is either part of the graph or it is not.

In this case I spoke of a set of points or of point sets. My warrant to do so derives from the use of real numbers to measure position along each axis and in the correspondence I discussed in Chapter 4 between the real numbers and points on a line.

To regard a line as consisting of mathematical points, in this sense, says nothing and assumes nothing about physical lines

this sense, says nothing and assumes nothing about physical lines or about other magnitudes that line segments might be used to represent. It does not imply or presuppose that a physical line is made up of points.

We encountered the same essential issue in Chapter 1. We observed, for example, that to treat a physical line as continuous is to say, in a particular context, that its possible discontinuities are immaterial. Similarly, this aspect of analytic geometry says nothing about metaphysics: its use is methodological.

One is interested in this set primarily because it has the shape of a circle.

Now consider the line 3y = -4x. The solution set to this equation is also infinite. But it intersects the circle in just two places, namely (x,y) = (-3,4) and (x,y) = (3,-4). One says that the intersection of the two solutions sets consists of these two points.

Describing point sets in this way is a natural consequence of the marriage between geometry and numbers (including algebra) known as analytic geometry. And, in this sense, set theory has roots in Descartes and Fermat.[9] In analytic geometry, geometric shapes are specified by algebraic equations, equations that measure certain characteristics of the shape. For example, the equation of a circle specifies and reflects the position of its center and its radius. To find an intersection of two geometric figures, the lifeblood of

Euclid’s

Elements, as discussed in Chapter 1, is to find the

simultaneous solutions of their corresponding equations.

As a final example of a mathematical set, suppose that one

wants to delimit a particular kind of number, say numbers divisible

by 5. Mathematicians would characterize the totality of qualifying

numbers as a set. Now consider a second set of numbers, those with

a remainder of 1 when divided by 5. If one chooses a member from

each set and adds them together, the total is always a number in the

second set consisting of numbers with a remainder of 1.

Studying remainders with respect to division, in this way, is

a step on another long journey, one that begins with the childhood

game of distinguishing odd numbers from even numbers. One

discovers that categories of numbers can have their own arithmetic:

a) even plus even or odd plus odd equals even and b) odd plus even

equals odd.

In every one of these examples, there is value in looking at

the discriminated range of mathematical objects as a set, as a

totality, but also as a totality of *individual members*.

In all of these cases, it is completely unambiguous whether

any particular concrete of the specified kind is included or not

included in the set. In the first example, 2 is a solution to the cubic

equation; 3 is not. In the fourth example (-3,4) is on the circle; (4,4)

is not. A number or a point is either included in a particular set of

numbers or, respectively, set of points or it is not.
Keep in mind, though, that *unambiguous* does not mean

obvious. To know that an equation has a solution set is not to have solved the equation. But to say that the set of solutions is

unambiguous *is* to say that the set of solutions, that whether a particular number would or would not satisfy the equation, is open

to unambiguous *discovery*. Unlike a concept, a set never has

borderline cases.[10]
These examples have a number of other things in common.

First, all of them isolate a range of possibilities, a set of solutions to

a particular equation, a range of possible values for a measurement,

or a range of possible numbers satisfying a particular condition.
Secondly, there is something about the particular members

of each set that makes them of interest; they are special in some

way in some particular context. Yet none of these sets corresponds

to a concept.
One does not, normally, for example, form a *concept* of the

solutions to a particular polynomial equation. Rather, the set of

possibilities is characterized, delimited, by some sort of description,

namely that they satisfy the equation. To solve the equation is to

find a more direct or, possibly, a more valuable characterization.
In the case of sin(s) = 0, the solution set is completely

characterized, completely specified, by the equation. Finding that

the solutions consist precisely of numbers of the form nπ (n an integer) is a discovery about something that has already been *specified* in some other way. Both the equation to be solved and the solution are characterizations; neither is a concept. To specify a set of solutions to an equation is to isolate things of a particular type that have something in common. But neither an equation nor its solution set normally functions as a permanent unit of thought in the way that a concept does.

Notice that to isolate a set requires conceptual means. If a set is finite, as in the first example, one can simply list its elements. This listing is, of course, done conceptually and, in the first example, it presupposes the concept of number. Still, in a mathematical context, listing the elements of a set is the most direct means of specification that is ever available. And it is only available for finite sets. More generally, *solving* an equation means finding a more direct specification of a set that has already been indirectly specified by the equation. A typical kind of example is the equation sin(s) = 0. One solves it by giving a formula for its solution set, namely, {nπ, where n is an integer}.

In saying that this formulation identifies the solution, one, first of all, connects the equation sin(s) = 0 to something else that one already knows, namely the irrational number π and the integers. But it is also implicit that both characterizations, both the

problem and the solution, describe the *same* set. The set that they describe has an ontological status, some form of existence that is

independent of *either* characterization. To speak of the solution as a *set* is to take precisely this perspective on the solution.

As one last observation, one *does* form a concept of

numbers divisible by 2; one says that such numbers are *even*. But one would not form a separate concept for divisibility by other

numbers, such as the number 5. A description is all that one needs.

One does something else: One finds a general way of expressing

such relationships. One says that $x = y \pmod 5$ to express the fact

that $(x – y)$ is divisible by 5. And, in particular, $x = 1 \pmod 5$ means

that $(x – 1)$ is divisible by 5, which is to say that x divided by 5

leaves a remainder of 1.

One does not, in any of these examples, form sets that mix

different kinds of things. One does not form a set consisting of both

numbers and circles. There is, in every case, a universe of discourse

or, as I prefer to call it, a mathematical domain: a demarcation of

instances of a valid, previously identified,

mathematical

abstraction.

These sets exist conceptually in the sense that, in each case,

one has isolated its members. But being included in the set does not

change the ontological status of its members in any way. The

members remain specific instances of the kind of thing that they

are. Numbers, for example, exist as part of a system of

measurements potentially applicable to a concrete. A *set* of

numbers

arises as the referents of a characterization, a

characterization that isolates numbers of particular interest in

some particular context. One *characterizes* the set, but the set itself is regarded simply as the numbers that have been isolated, qua

something that has been *isolated*. A different characterization that happened to *isolate* precisely the same individuals might reflect a completely different conceptual perspective, but it would,

nonetheless, be a *different characterization* of *the same set*. A set and its members are not distinguished by *how* the members have been isolated, but only by the fact that these members *have been*

isolated.
In this respect, membership in a set is very different from

being a referent of a concept. There is always a mathematical

concept, such as real numbers, for the *kind of thing* that one includes in a particular set. But sets, as such, are not concepts, even

though one uses conceptual means to isolate them. To be a referent

of a concept presupposes a particular perspective, that one

recognizes an axis of *similarity* among the referents. But to be a member of a set only requires that it be demarcated or isolated

somehow, as belonging to the set. The precise *how* has no bearing, is like an omitted measurement.[11]
This last sentence requires some explaining. For, I said that

a set is not a concept. It is, rather, a conceptual isolation of

instances of a concept. And Ayn Rand's observations concerning

omitted measurements are intended to apply specifically to

concepts. Yet, in a broader sense, the same principle applies to a

set: The possibilities isolated must be isolated by some conceptual

characteristic. But the specific possibilities that are isolated do not

depend on the particular means by which they are isolated.
For example, the set of points that satisfy $x^2 + y^2 = 25$ is the set of points that fall on a circle of radius 5, centered at the origin

(0, 0). Notice: two specifications; one circle, one set of points.
There is indeed a concept involved whose measurements

are being omitted: the concept of isolation.

Isolation of the

members of a set is independent of the specific details regarding

how that isolation was accomplished. Of each element, one says

only that it has been isolated, has been included in the set. *How* it was isolated doesn't matter.
If the notion of a set in *mathematics*, if the notion, say, of a

set of numbers, is to have value in mathematics, one must embrace

the possibility, for example, of an *infinite* set of numbers. Neither numbers nor sets of numbers exist in the world as such. But

numbers

do exist

as

identifications of relationships, as

measurements. Numbers are a system of concepts, indeed, of

measurements. The *relationship* that a number measures is

categorical. It is independent of any specific concrete to which it

applies. Treating any *specific* application as an omitted

measurement within a mathematical pursuit, one can treat

numbers and sets of numbers as objects of investigation, as existing prior to or independent of one's investigation. *Not existing as a completed infinity, but as an isolated specific demarcation of*

potential measurements of magnitudes. Any number that one

isolates is a particular measurement that might be applied to a

concrete or involved in a calculation.

Sets, at root, are a methodological device to isolate and

keep track of distinguishable mathematical possibilities of a certain

kind. To consider an element or selected elements of a set is to

isolate and consider a range of possibilities. To look at an isolated

range of mathematical possibilities as a set is to take a geometric

perspective on that range of possibilities. It is to look at the

members of the set as, in some sense, *objects* of thought.

A mathematical set is infinite when the distinguishable

mathematical possibilities are unlimited. For example, the real

numbers are unlimited in that the concept of real number

recognizes no prior limitation regarding either potential precision

or multiplicity. At that level of abstraction the concept of real

number distinguishes an unlimited range of possibilities.

Mathematical sets do have limits, but those limits are imposed, first

of all, by the genus (e.g., real numbers), by the kind of thing

included in the set and, second, by whatever conceptual means is

applied to isolate their members.

A concept or a class is not, as such, a set. *Man* is openended in a sense that *number* is not. Numbers form a conceptual system and they are tightly circumscribed in the ways that

individual numbers can differ from each other and relate to each

other. One *grasps* the specific mathematical domain by grasping, as part of a single dimension of variability the *spectrum* of possible numerical relationships to a unit.

One does not, in similar fashion, grasp men as, in any

sense, a domain. One is, first of all, aware of a vast array of respects

in which one man can differ from another. Secondly, it would be

unreasonable to presume that there are not yet other axes of

variation that have not yet been discovered. Finally, men, as such,

are concrete individuals.

And this is another point of contrast: A number is a

particular quantitative relationship that transcends any particular

instance. A mathematical set *never* consists of concrete entities or concrete instances of a first-level abstraction.

In regard to dimensions of variability, take *color* as an

example somewhere in between *man* and *number*. Color is also multi-faceted. But one can focus on a single dimension of color,

such as hue or intensity. Or one can even focus on a specific

constellation of dimensions, such as hue, intensity, and saturation.
If color varies along still other dimensions, presently

unidentified, one can regard such other dimensions as omitted

measurements. And if there should later turn out to be previously

unknown facets of, say, hue, then one's focus on the particular

facets that one *has* identified implicitly treats any such additional facets that might be discovered *also* as omitted measurements, as irrelevant to the distinctions one is making. In sum, this is to say

that one regards any two colors that have the same hue, intensity,

and saturation, as the same color, treating any other respects in

which they may differ as irrelevant in the particular context.
Nonetheless, with all that said, isolating a dimension, such

as hue, is not yet to completely grasp the mathematical

relationships between hues; it is not to grasp the precise respect in

which one hue differs from another hue, nor is it to grasp the full

potential variations of hue.
But one is getting closer. In grasping that colors differ by

hue, intensity, and saturation, one *circumscribes* an area of study in a way that parallels the specialization of a mathematical study.

One singles out a specific constellation of characteristics that differ

from each other along specific, known, measurable, dimensions. In

focusing on hue, one isolates and orders the visible spectrum.
So how is mathematics different from other

specializations? When one considers

measurements, such as

numbers, then what one isolates is delimited in a way that “man” is

not. This isolation is openended, but, speaking very loosely, not

wide openended.
The

application of number *is* wide openended.

Applications of mathematics to the world have no known limits

regarding either the concretes to which they might apply or the

manner of that application. *Applications* of numbers are not, in Cantor’s phrase, “welldistinguished objects.”[12]
But, by contrast, the range of possible numerical

measurements is isolated as a *system* as described in Chapter 4.

Number is graspable as a specific *dimension* along which a quantity that it measures can vary or can relate to a unit. One has a *specific* grasp of the delimited *respects* in which two numbers can differ

and of the range of possible measurements. Indeed, it is the job of

mathematics to provide such a grasp, to provide ways to cover the

full range of possibilities, to provide a comprehensive *system* of measurement.
What about magnitude? *Magnitude* is openended in a way

that *number* is not. Although all magnitudes have certain things in common, there is an apparently limitless variety of kinds of

magnitudes, a sampling of which I reviewed in Chapter 2. And for

each kind of magnitude there is an openended range of entities

possessing attributes of that type. For example, the range of objects

possessing attributes of that type. For example, the range of objects

possessing length is openended in exactly the same way that the

concept *man* is openended.
But suppose one considers magnitudes solely in regard to

the *relationships* between them, between two magnitudes of the same kind. Suppose that one looks at them as measurable along a

single dimension. Suppose a perspective from which any

consideration of a concrete case embodies an *abstract perspective* applying to the entire category of measurements, of magnitudes as

such, the kind of perspective that I applied, in Chapter 1, to the

drawings in Euclid's *Elements*. Suppose, in short, that one looks at magnitude *geometrically*, as a continuum of related possibilities, as I did in Chapter 2. Then there is only one *relevant* difference between any two magnitudes of the same kind. And that difference

is given by the numerical ratio between them.
To grasp continuous magnitudes of a particular kind, one

needs a concrete example. And to understand the way that two

magnitudes of particular kind can differ, qua magnitude, requires

scientific discovery. But once that discovery has been made, one

grasps the entire range of possibilities for that particular

characteristic along the particular dimension that one has

identified. One has grasped the respect in which any magnitude of

that type relates to a unit. That range of possibilities is captured in

the real number line.
As presented here, a set requires the ability to establish

systems of measurements and, in geometric contexts, to isolate a specific constellation of measurable dimensions. A set is formed in reference to a mathematical abstraction and requires an unambiguous demarcation among the units of that abstraction; it requires a demarcation of specific possibilities distinguished by that abstraction. Finally, a set involves a specific focus, on the possibilities or contingencies thus isolated, as objects of investigation, considered without regard to *how* they were isolated.

The intent of this chapter is to develop a reality-based approach to mathematical set theory without reference to any particular axiom system. Later in this chapter, I will discuss the early history of set theory, culminating in the Zermelo-Fraenkel (ZF) axioms.

For now, it must be recognized, that *something* must limit the domain of set theory. There is no such thing as a “Set: set of all sets.” That alleged concept presupposed some kind of pre-existing conceptual universe and was abandoned when it led to contradictions.

The modern answer to the problem that these contradictions exemplified consists in laying out axiom systems, such as the ZF axioms. But in this chapter I take a very different approach. I take a much more openended, inductive and hierarchical approach, one that builds from the ground up. Reality

hierarchical approach, one that builds from the ground up. Reality is the foundation, the ultimate point of reference, and all of our mathematical concepts, no matter how abstract, are formed hierarchically, building on earlier concepts tracing ultimately back to our direct observations about the world. To understand something one must, first, relate it to the world. Tracing relationships among ideas, organizing one's knowledge, critical though it be, is not a substitute for understanding, for connecting it to the world.

If, in such an approach, contradictions are discovered along the way, the fault lies in some error of identification. Reality and our cognitive needs set the ultimate standard and the ultimate point of reference. The world does not contain contradictions. If one is discovered, the fault is not in the world, but in us, in our misidentifications. Turning one's back on the world is not a solution to the problem; it is an evasion of it.

In general, I reject an approach that, in the manner of settheoretic axiom systems, attempts to lay a deductive foundation and circumscribe the entire domain of mathematics in advance, in the hopes that nothing more will ever be needed and that no contradictions will ever emerge. The suitability of settheoretic methods within a particular mathematical abstraction is something that, in any approach, must be discovered and cannot be presumed

in advance.
My approach to understanding set theory begins with the

concept of a mathematical domain, the subject of the next section.

Mathematical Domains

Measurement involves relationships among things that are

similar in some way. A measurement determines a respect in which

two similar things are different, in which two things differ along an

axis of similarity. If the object of measurement is a *magnitude*, the expression, the *measurement,* of this relationship is a *number*. A *measurement,* in general, is an expression of a relationship to a unit, an expression of how the attribute or existent being measured

differs from the unit in a particular respect.

If one takes an abstract perspective on the

object of

measurement, as an object of measurement in a

particular

respect(s), then one's focus is *geometric*. A domain of geometric possibilities is a range of possibilities with regard to a specific

constellation of measurements, possibilities that are distinguished,

as relevantly different, only with respect to those measurements.

If one focuses on the

relationship of the characteristic one

is measuring to its unit, if one focuses on the *measurement,* one's focus is

arithmetic or algebraic. Within either

perspective,

geometric or arithmetic, the object of one's focus is a *mathematical domain*. The real number

line is a *geometric domain*, a

demarcation of possible magnitudes of any type whatever,

considered in relation to a unit of the same type; the real number

system is an arithmetic domain, a *domain of measurements* applicable to magnitudes.

I use the term, and concept,

mathematical domain to

replace the standard settheoretic term, *universe*. In my usage, a *mathematical domain* consists of a demarcation of instances of a valid, previously identified mathematical abstraction, of

mathematical possibilities of a particular kind, considered as an

object of investigation. I prefer the term *domain* to *universe* because I hold that the concept of a mathematical *set* only makes sense within the context of a mathematical *domain* in the sense I have just characterized. Generally speaking, a mathematical

domain can also be considered a set, but it is a domain *first* and a set *second*. A mathematical domain functions as a universe of

discourse. But *mathematical domain*, as a concept, is a distinct concept from the notion of *universe of discourse*, and represents a different conceptual perspective.

The qualification of a mathematical domain as a

set presupposes and requires a level of mathematical abstraction that

regards *specific* external referents, such as the type of magnitude measured by a number, as omitted measurements. A number is a

relationship to a unit; everything else is unspecified and the mathematical relationships among numbers do not depend upon

the particular concretes to which they might be applied, the

concretes subsumed by the abstraction. Numbers must be

applicable to concretes and relationships between numbers reflect

and subsume relationships among concretes, but what those

concretes might be is an omitted measurement.

In the case of a domain of geometric possibilities, one

regards these possibilities as related by a specific constellation of

measurements and conceptually distinguished only with respect to

those

particular measurements. For example, in the case of

magnitudes, one demands that the scope of possibilities, in any

particular case, all represent the same kind of magnitude and treats

them as distinguished only by their differing relationships to a

chosen unit. The type of magnitude involved and the choice of

specific unit are omitted measurements.

As I mentioned earlier, the mathematical concept of a set is

a concept of method. One isolates sets in an abstract mathematical

context in which concrete referents and, in particular, any limitations to

precision are regarded as omitted measurements.[13]

Consider, for example, a converging infinite set of points,

1.9, 1.99, 1.999, …

as it might relate to the centimeter markings on a meter stick. As a mathematical sequence, every number is, unambiguously, either in

the sequence or not in the sequence. 1.9999 is in the sequence; 1.95

is not. However, on any particular meter stick only a very small

number of these numbers can actually be distinguished. The points

on the meter stick have a limit point, namely 2. But most of the

points in the sequence are indistinguishable from that limit point.

And the placement of all these points, both the finite number that

can be distinguished and those at the limit point that cannot be

distinguished, are subject to the particular standard of precision

that one achieves in the particular case. In effect, the sequence on

that particular meter stick is 1.9, 1.99, 2, 2, 2 … .

So the concept of an infinite set does not apply, per se, to a

prescribed sequence of markings on a meter stick. As I remarked

earlier, there are no borderline cases regarding membership in a

set. Membership must be unambiguous. But this condition, in the

case of infinite sets, can only be realized, as a methodological

device, within the scope of a specifically mathematical abstraction.

As applied to any specific concrete, there are always a finite set of points and their placement is subject to a specific precision limit. The concept of an infinite set is a creature of mathematics and refers to a range of mathematical possibilities or contingencies. Numbers are useful, and perhaps even necessary, in building or specifying infinite sets precisely because they provide a limitless, even indenumerable system of measurements, unambiguously distinguished and specifiable. To qualify a set, one needs a way to unambiguously distinguish each member and unambiguously specify, for each, which elements of the broader universe are in the set. Thus, one can specify the set of even numbers because any natural number one might name is either divisible by two or it isn't.

This, of course, is not the conception of a set as originally advanced by Cantor. As we shall see later, however, Cantor's early, naïve perspective on sets was, famously, soon abandoned. So, to a significant extent, to state my negative viewpoint on Cantor's conception is to beat a dead horse. But I state it anyway: Regarding the scope of the concept *set*, it is simply incoherent, for example, to characterize the aggregate of all distinguishable objects in the world, combined, say, with all potential abstractions, as a *set*. There is simply no end to it, no specific limits, no unambiguous

distinctions, and no cognitive purpose to be served. Neither the world nor our conceptions of it come to us carved up into discrete units. Anything in the world can potentially be counted or distinguished: objects, attributes, measurements of magnitudes, distinguishable parts of things, with other parts, ad infinitum, cutting across those parts, relationships between things, relationships between relationships, second thoughts, vague emotions, *etc.* And there is no reason on earth, or in mathematics, why anyone would ever need such an aggregation of disparate units. Meaningful use of the concept *set* requires specific isolation, clear, unambiguous, distinction among its members and a wider conceptual category within which these members are distinguished.

I hold that the concept of an infinite set is applicable only in a mathematical context in which any *specific* reference to realworld concretes is treated as an omitted measurement. In its systematic scientific use, sets are a mathematical concept, a methodological device, applicable specifically to mathematical abstractions, as such. For a domain to qualify as a set, its elements must be unambiguously differentiated and, yet, interrelated, as part of a system, in ways that specifically delimit its scope.

The concept of a *mathematical domain*, consisting of a system of measurements or geometric objects of some kind, has

system of measurements of geometric objects of some kind, has

fundamental importance in mathematics. I introduced this concept,

somewhat informally, in Chapter 4 and I characterized the domains

of natural numbers, integers, rational numbers and real numbers.

The concept of a

mathematical

domain, as applied to

measurements, is really just a different perspective on a *system* of measurements, one that focuses on, as particulars, the

measurements that are embraced and distinguished in the system rather than the *system* itself. But it focuses on those measurements as constituting particulars within that *system* of measurements.

When one analyzes relationships between measurements, such as

the laws of addition in the case of numbers, one is treating these

measurements as sharing a domain. The laws of addition apply

equally to all numbers within the number domain.

My usage of the term,

domain, is not standard though it is

closely related to a standard use of this term. So, to avoid

confusion, a brief digression is unavoidable.

The closest

standard equivalent to my

concept of

mathematical domain is the settheoretic informal term *universe of discourse*. In my usage, a mathematical domain is indeed a

universe of discourse. However, it is so, not by fiat, but by virtue of

arising as a system of measurements, arising in *relation* to a system of measurements, or arising as a geometric object, an object of

measurement considered in relation and only in relation to a

particular constellation of measurable characteristics.

The

standard use of the word

domain applies to a

mathematical function (i.e., a relationship between an independent

variable and a dependent variable). One speaks of the *domain* of a function, thinking of it as the set on which the function is defined. I

will use the term *domain* in this way, as well. When the context is unclear, I will use the full term *mathematical domain* as the rough equivalent of mathematical universe of discourse, in the one case,

and speak of the *domain of a function* in the second case.

Analytic Geometry as a Mathematical

Domain

Mathematical domains are identified and specified

conceptually and they are related hierarchically. I have spent most

of my time in this book examining the roots of mathematical

concepts in reality, in the measurement of the world. But, in

concepts in reality, in the measurement of the world. But, in mathematics, one thing always leads to another. A new mathematical or scientific problem requires and leads to new insights, new integrations, new connections, new concepts, and, sometimes, new mathematical domains, entirely new fields of investigation.

Analytic geometry, the integration of number and geometry, of algebra and geometry, was the first decisive step in modern times beyond the geometry of the ancient Greeks. Analytic geometry was discovered independently, and at about the same time, by Descartes and Fermat in the 17th century.[14]

The classical Greeks had used line segments to represent both magnitudes and multitudes and they had reasoned geometrically about both, but the idea of the number line, of identifying numbers with points on a line, had never occurred to them. Indeed, the classical Greek's conception of ratio as a relationship between two quantities of the same kind and, therefore (from a modern perspective) dimensionless, severely inhibited such discovery. One cannot compare a *ratio* of two lengths, each expressed, say, in feet, with a *magnitude* such as length. A ratio is not measured in feet and is, in fact, independent of any unit.

And this distinction, and awareness of this distinction, was

reflected in Greek practice. Euclid used lines to represent magnitudes and multitudes, but never to represent ratios. Rather, a ratio was *always* represented by a *pair* of magnitudes or multitudes. Euclid understood the distinction between ratio and magnitude. He was unable to integrate them into a single system of measurements.

Descartes' later discovery of analytic geometry was, in part and quite explicitly, a revolt against this aspect of the classical approach to quantity. The Greeks could apply geometric reasoning to non-geometric contexts, but they could not reason in the opposite direction. And, of course, the Greeks had not discovered algebra. As a consequence, they could produce sophisticated abstract geometric arguments, but were, for example, completely dependent upon a geometric perspective to make abstract arguments pertaining to numbers.[15]

As another example, the Greeks studied conic sections, but their study was enormously complicated by their reliance on classical geometric methods. To perform an abstract measurement, in ancient Greece, always meant to construct, or at least produce, a

suitable line segment to represent a linear magnitude, a twodimensional figure to represent an area, a cube to represent a volume, or a *pair* of magnitudes of some kind to represent a ratio. The classical treatise on conic sections by Apollonius[16] abounds with just such constructions, with arguments and conclusions that

are difficult to follow, to retain, and to appreciate.

Analytic geometry married algebra and geometry, opening

geometry to algebraic methods and creating a passageway between

two hitherto independent and unconnected disciplines. And the

immediate consequence was to vastly simplify the study of conic

sections and of other shapes that were studied in antiquity.

As a mathematical domain, analytic geometry identifies

each axis with the real number line and each point in the plane with

an ordered pair of real numbers, with the first coordinate

representing the *x* axis and the second coordinate representing the *y* axis. Analytic geometry, the Cartesian coordinate system,

depends hierarchically on the real number line.

Cartesian coordinates provide the theater in which one

relates quantitative relationships between the coordinates to the

shapes that these relationships capture, as embodied in their

graphs. Expressed in Cartesian coordinates, a graph is a set of

points satisfying a particular relationship.

For example, $x^2 + y^2 = 25$ is the equation of a circle consisting of points 5 units away from the origin (0,0). The point x

= 3 and y = 4, represented in Cartesian coordinates as (3,4), lies on the circle. It lies on the circle because it satisfies the equation. That is, $3^2 + 4^2 = 9 + 16 = 25$.

Prior to the introduction of analytic geometry, one did not think of geometric shapes as consisting of points; one followed Aristotle in thinking of them as wholes that were divisible into smaller wholes. One could intersect two lines at a point or even *choose* a point of division on a line, but a point of division was just that, a *division*, not a part of the line. To be part of a line was to be divisible.[17]

Analytic geometry fundamentally changed that perspective. Within a Cartesian plane, coordinates became the universal means of measurement of any geometric figure. A geometric shape was characterized as the set of coordinate pairs, of points, satisfying an algebraic equation relating the coordinates, as in my circle example. One no longer resorted to constructions, the way Euclid had, to compare one thing with another: Every point on the Cartesian plane came equipped with coordinates that measured its position with respect to the two axes. A pair of numbers determined a point; every point had two coordinates that measured its position. Every relationship in the Cartesian plane was now, directly or indirectly, to be described in relation to a coordinate system. To specify a shape, one specified the coordinates of the points included in the

shape, one specified the coordinates of the points included in the shape. To relate two points was to relate their coordinates. The coordinates, the means of measurement, became the immediate object of further measurement. Point sets, at least implicitly, had entered mathematics.

In a very primitive sense, especially as applied to measurements or to points, to specify a set is to make a measurement. A set is a demarcation. A set of measurements, such as numbers, isolates a range of measurements. A set of points in analytic geometry isolates positions in an abstract plane, positions that are identified or identifiable by coordinate pairs.

A set is *not* a measurement in the full sense, even in the abstract sense I discussed in Chapter 1 and elsewhere. For an abstract measurement identifies a *quantitative relationship*. By contrast, to characterize a set is *not* to identify a quantitative relationship; it is only to *distinguish* certain instances of an existing category of measurement.

Nonetheless, on the most primitive level, one function of measurement is to *make distinctions* among similar concretes. On the one side, a *concept* identifies a *similarity* among differences.

On the other, a *measurement* identifies a *difference* within a similarity. Within an abstract mathematical setting, as applied to a mathematical abstraction, a set does exactly that in the most

primitive terms. It identifies a difference within a similarity. A measurement distinguishes by relating, by establishing a quantitative relationship. A set does not relate, but it does distinguish.

The function of isolation is needed, in just this way, in a mathematical context. A set performs the primitive function *of isolation*, of distinguishing something that one is interested in from a broader category of similar existents within a mathematical domain. To provide a point of reference and to maintain a connection to the world, a set requires a genus, a specific broader category that it differentiates, just as a definition does. In the deepest sense, this is why a meaningful concept of set presupposes a wider abstraction; i.e., it presupposes a genus. The required genus is supplied by that wider abstraction.

Geometrically, a circle is a shape. But analytic geometry measures a circle as a set of points. One specifies a circle either as a set of coordinates or as an equation in numbers whose solution is a set of coordinates. And one can move in both directions. One can start with a circle, described, say, in reference to the coordinates of its center and a given radius, and, from that, find the equation that describes the circle. Or one can start with the equation and

characterize its solution set in some way, by some other means. In

either case, one *specifies* a set of coordinates, but by different means.

One looks at the Cartesian plane geometrically, but it is a

measured geometry.
Normally, one thinks of a Cartesian plane as being given by

its coordinate system. But one also speaks of

changes in

coordinates. To change coordinates is to regard the Cartesian plane

as fixed, as referring to something external, but as being measured

by a different coordinate system, a different pair of axes. One

thinks, for example, of rotating the *coordinate system* and this action is different, conceptually, from rotating *things* in the plane against a fixed coordinate system. Since the rotated coordinate system is simply a different set

of axes, one can represent the *new* set of axes by two perpendicular lines through the origin, the (0, 0) point in the original coordinate

system. Any point on the Cartesian plane can now be identified in

two different ways, each way corresponding to one of the two

different coordinate systems. From this perspective, one looks at

the points in the Cartesian plane as being *specified* in two different ways and having a kind of existence independent of either

coordinate system, by virtue of which one can relate the set of

measurements, the coordinates, for the first coordinate system to the coordinates for the second coordinate system. In developing

this relationship one necessarily thinks geometrically of something

external that is being measured. One thinks of something that

exists, independent of the particular means of measurement.
In regards to the plane, one implicitly treats the specific

means, the particular coordinate system, of measuring it as an omitted measurement, a measurement that one is now in the

process of supplying. Yet because the plane is an abstraction, not

actually tied, in a *mathematical* context, to any *particular* concrete, one can only *specify* a set of axes with respect to some other description of the plane, such as the description provided by the

original set of coordinates.
And the issue, and the limitation, here is epistemological. It

is right to treat the plane as concrete because, whenever one *does* apply the mathematics to objects in the world, the coordinates in

the plane measure those *objects in the world*: In *any* application, the Cartesian coordinates are specified in relation to these objects.

In describing a change in coordinates, one specifies a way to change

one concrete specification of coordinates relating to an external object to *another* concrete specification of coordinates relating to that same object. That relationship *between the coordinate systems* is entirely mathematical simply because that relationship is

independent of any *specific* object to which it might apply. The relationship between coordinates is an abstraction that applies

equally, and in the same way, to all of its referents.
When one looks at the Cartesian plane as externally given,

this, the geometric perspective, is the perspective that one is

actually taking. The use of a first coordinate system as a point of

reference to specify a second coordinate system does not change the

essence of what one is doing with respect to external reality.
Analytic geometry also provided a new perspective on

numbers, regarded as a system of measurements. To identify

numbers with points on a line is to externalize numbers, to bring

them into focus as objects of investigation, as part of an extended

entity with a determinate structure having an independent

existence as a system of abstractions. A number is regarded as part

of a totality and in relation to that totality.
Numbers are, of course, related arithmetically. For

example, 2 + 3 = 5 expresses an arithmetic relationship. But they

are also related geometrically: the number 2 is closer to 5 than it is

to 7. The number line emphasizes this more geometric perspective.

Numbers are *not* abstract magnitudes, but they

measure magnitudes and the relationships between numbers reflect and,

indeed, *pertain* to relationships between magnitudes.
But more fundamental than the *visualization* of the

number system is the isolation and treatment of the number system

as an *object of study*. Number, a means of measurement, thereby (but derivatively)

becomes, at one remove,

an

object of

measurement

measurement.
And why study number, as such? For a now familiar

reason: One establishes relationships between numbers to facilitate indirect measurement. The ability to add and multiply reduces

one's need to count, enhancing and facilitating one's grasp of the numerical relationships in the world. Relationships between

numbers are, at root, relationships between the things one uses

those numbers to measure. Numbers matter because the things that

numbers measure matter.
So, in treating numbers as an object of study, as a domain

of discourse, one is not inventing, creating, constructing,

discovering, or intuiting a mathematical universe. One is simply

adopting a conceptual perspective on *this* universe, a universe that exists independently of one's knowledge of it. One is dealing with

relationships that actually exist in the universe and developing

methods to deal with such relationships in general. Numbers are an

object of discovery because the reality that they capture is an object

of discovery. Every relationship between numbers captures

relationships among measurements that exist or that might exist,

that relate, or, in the nature of things, *might* relate actual existents in the world that we live in.

Mathematical Domains as Hierarchical

The study of polynomial equations, such as $x^3 - 6x^2 + 3x +$

$10 = 0$ predates the development of algebra. Albeit in numerical

10 = 0 predates the development of algebra. Albeit in numerical

form, methods for solving quadratic equations (such as $x^2 - 3x - 4 =$

0) were discovered in antiquity.[18] But the invention of algebra led, in modern times, to explicit solutions, in terms of radicals (square

roots, cube roots, etc.) for cubic and quartic (degree 4) polynomials.

Beyond the quest for explicit solutions, mathematicians also sought

techniques to find approximate solutions of polynomial

equations.[19] But to develop effective approximation techniques to solve polynomial equations, one needs to know something about

the general shape of the graphs of the polynomials.

So polynomials became a specialization, a study once called

the *theory of equations*, and incorporated today into the broader study known as *algebraic geometry*. Polynomials became a

mathematical domain of investigation.

Polynomials represent measurements in a variety of

respects. Most importantly, as a function, as a relationship between

x and y, a polynomial equation such as $y = x^3 - 6x^2 + 3x + 10$

specifies a potential relationship between two magnitudes.

Secondly, in relation to its graph, a polynomial equation represents

a condensed numerical expression of a geometric shape, capturing

and specifying it in one equation. And, finally, to set the polynomial

expression to zero (e.g., $x^3 - 6x^2 + 3x + 10 = 0$), is to *indirectly* specify a solution set, the *roots* of the equation. One solves the equation to identify these roots

explicitly, but the equation *specifies* a set of numbers simply by virtue of the fact that solving the

equation is an act of *discovery*, a discovery of something that has *already been specified*.

So a polynomial

relates to measurement in a variety of

ways. But, in the first respect, as a function, as a quantitative

specification of a relationship between two variables, it also *counts* as a measurement.

Cubic polynomials, such as $y = f(x) = x^3 - 6x^2 + 3x + 10$, as a species, are a *system* of measurements. They have their own

arithmetic. For example, one adds two polynomials, by adding their

y values, point by point, for each value of *x*. If $y = g(x) = x^3 + x^2 + x + 2$

is a second cubic polynomial, one obtains their sum $y = (f +$

$g)(x)$ as $(f + g)(x) = f(x) + g(x)$.

One can make this more concrete in two ways. First, if $x =$

1, then the value of $f(1)$ is 8 (by substitution for x) and the value of

$g(1)$ is 5. So the value of $(f + g)(1)$ is $8 + 5 = 13$. Secondly, since for

any particular value of x, both polynomials are numbers, one can

add them term by term. One finds

$$(f + g)(x) = f(x) + g(x)$$
$$= (x^3 - 6x^2 + 3x + 10) + (x^3 + x^2 + x + 2) = 2x^3 - 5x^2 + 4x + 12.$$

If one substitutes a value $x = 1$ in the right-most expression,

one checks that the value of (f + g)(1), of $2x^3 - 5x^2 + 4x + 12$, where x = 1, is 13.

Polynomials are related to numbers in at least two respects.

Most importantly and fundamentally, if one substitutes a number

for x into a polynomial, one gets a number. Secondly, one requires

four numbers, specifically four *coefficients*, to *specify* a particular cubic polynomial. The polynomial *f*, for example, has coefficients 1

(implicitly),

-6, 3, and 10. The second polynomial,

g, has

coefficients 1, 1, 1, and 2. As a system of measurements, cubic

polynomials differ from each other in specific, circumscribed,

quantifiable ways. In isolating cubic polynomials as a realm of

mathematical study, one *grasps and isolates the specific*

dimensions along which two cubic polynomials can differ from

each other. Cubic polynomials constitute a mathematical domain.

Note that to

isolate a particular mathematical domain, as

with concepts generally, is not to individually specify or “Construct”

each referent covered by the concept. As with numbers, one does

not have to specify or construct each particular member of the

domain in order to implicitly include it. Ayn Rand’s analogy to

express the openended character of concepts is relevant here:

express the open-ended character of concepts is relevant here:

“An arithmetical sequence extends into infinity,
without implying that infinity actually exists; such
extension means only that whatever number of
units does exist, it is to be included in the same
sequence. The same principle applies to concepts:
the concept “man” does not (and need not) specify
what number of men will ultimately have existed –
it specifies only the characteristics of man, and
means that any number of entities possessing
these characteristics is to be identified as
“men.””[20]

Any demarcation of certain cubic polynomials from the domain of cubic polynomials is a set. For example, cubic polynomials characterized as having the first coefficient equal to one, constitute a set of polynomials. And this, for two reasons: First, one can speak, in general, of a specific set of cubic polynomials, because one has completely isolated the mathematical dimensions that distinguish one cubic polynomial from another and identified the range of possibilities. Cubic polynomials are a set. Secondly, because, in this particular case, any cubic polynomial

one might consider, either, unambiguously, has its first coefficient equal to one or it does not.

To give a second example, to which these same considerations apply, there is a set of cubic polynomials that vanish, that are zero, when x = 1. Again, this is an unambiguous further specification within the set of cubic polynomials

More generally, any unambiguous specification that either applies or does not apply to each cubic polynomial specifies a *set* of cubic polynomials. This includes the entire domain of cubic polynomials, considered as a totality: Again, the mathematical domain of cubic polynomials can be considered as a set.

And, finally, within the context of, and with respect to, a mathematical domain, the socalled empty set that has no members can be counted as a set for the same reasons that zero is counted as a number.

I have found neither warrant nor purpose for taking unions and intersections between sets contained within different domains. I maintain that there is no such warrant, except insofar as one mathematical domain can be regarded as a sub-domain of a larger domain or insofar as there is a mathematical domain that contains both. To distinguish is to distinguish in a particular respect, as I pointed out earlier. To compare sets as to membership requires a

common genus. But please notice that, in weighing in against

unions and intersections across domains I do *not* reject Cartesian products from distinct domains or functional relationships between different domains.[21]

If a set is always a specification within a mathematical

domain, then it is false to say, with established set theory, that there

is a unique empty set. The empty set is a *relative* concept

pertaining, always, to a particular mathematical domain. There is a colloquial view that "there is nothing is in my pocket," presupposes a particular *kind of thing* that is, indeed, absent. To say that there is nothing in my pocket is not to say that my pocket contains a perfect vacuum.

In mathematics, zero is a particular number within a

specific continuum as part of the number system. The same principle applies to the empty set. To say I have no apples is

different from saying I have no oranges.

Now, it is true that a set is determined specifically by its

membership. But sets arise in a specifically mathematical context

and are only meaningful within a such a context.

Cubic polynomials are a mathematical domain, but their

definition, as such, presupposes the domain of number. And to

recognize cubic polynomials as a domain, one must first have

recognized the number system as a mathematical domain. The

abstraction of cubic polynomial presupposes the abstraction of

number. To identify the dimensions along which cubic polynomials

can differ and to grasp the full range of such potential differences, one must have already done the same for the number system. Typically, mathematical domains arise in the context of, and presuppose, previously identified domains, such as the number system. But not always: For example, there is a particularly important class of mathematical domains, abstract groups, that does not arise in this way. Partly for this reason, I discuss abstract groups in Chapter 8.

In general, to establish a system of measurements or a geometric structure as a mathematical domain requires either relating it to domains that have already been identified, or relating it directly to reality, discovering what it measures, how it measures, and delimiting a system of measurements along a specified constellation of dimensions, identifying specific axes along which two similar things can differ.

Functions

The discovery of calculus ultimately brought many other mathematical disciplines in its wake. Newton invented integral and differential calculus as a means to both the formulation and the solution of problems in physics. He needed both integral and differential calculus. For example, Newton’s discovery of universal

gravitation required him to show that the gravitational action of the earth on exterior bodies would be the same if all of its mass were concentrated at the center of the earth.[22] For this, he used *integral* calculus. Newton needed *differential* calculus to measure change, both with respect to time and with respect to position. The equations of motion require solving equations involving derivatives that measure rates of change, to discover the trajectory of the moving objects.

The task of solving differential equations, equations of various degrees of complexity, has been a central mathematical problem, a problem in indirect measurement, ever since. Although much is known about solving differential equations, an enormous amount is yet to be discovered. In special circumstances there are techniques for finding explicit solutions in terms of known functions. But, in general, effectively solving differential equations requires approximation techniques. And here, there is a complication: A solution to a differential equation is not a number; it is a *function*. What does it even mean to approximate a function?

As I discussed in Chapters 2 and 4, the most fundamental principles of approximation already arise in the study of real numbers. And, prior to that, we saw in Chapter 1 that understanding the proper role of approximation, of a standard of

precision, is one key to understanding Euclidean geometry. Any context to which mathematics applies, involves a specific degree of precision, a standard of perfection to distinguish the important from the irrelevant. The role of mathematics is to provide a general way to meet that, or any, standard, to provide a *method* that applies independently of whatever that standard might be and whatever its nature might be.

All techniques of approximation, even the various specifications of what approximation means, are reducible to numerical approximation. But there are a multitude of *respects* in which two functions can be close to each other and which respect is important depends on the context. I will elaborate this point further, in my discussion of point set topology, later in this chapter.

However, my immediate interest relates to mathematical domain: In order to approximate functions with other functions one needs, first of all, to relate functions to each other. If a function is a *measurement*, in the way that a polynomial is a *measurement*, then one needs to isolate functions, in general, as a system of measurements. One needs to know, generally, how to relate functions to each other before one can define what it means for functions to be close to each other in some respect. One needs to regard a class of functions of some particular type as a mathematical domain.

mathematical domain.

Consider, for example, continuously differentiable realvalued functions on the interval from 0 to 1. Informally, to say that a function is continuously differentiable means that the function

has a derivative at every point and its derivative is continuous,

meaning that the derivative does not suddenly jump from one value

to another value, that the direction of the tangent to the graph of

the function does not abruptly change direction at some point.

The sum of two functions is defined the same way that one

adds polynomials: if f and g are functions, then the value of the

function (f + g) at a point x is defined by the equation

$(f + g)(x) = f(x) + g(x)$

If f and g are continuously differentiable, then the function

(f + g) is continuously differentiable. One knows, from elementary calculus, that its derivative is the sum of the derivatives of the two functions.

In a similar fashion, one can define multiplication of a

function by a number. If A is a number, then multiplying a function

f by A means to multiply the value of f at every point by A.

Symbolically, the function *Af* is defined in terms of the function *f*, for each point x, by
$(Af)(x) = A(f(x))$

For example, if $f(3) = 5$ and $A = 7$ then $(Af)(3) = 7 \times f(3) = 7$

$\times 5 = 35$.
If f is continuously differentiable, then Af is also

continuously differentiable.

Continuously differentiable functions on an interval are *measurements*: They measure, they specify quantitatively, a particular relationship between two measurements, a relationship between the variable x and the variable represented by the value of the function. Just as numbers specify a potential relationship between two multitudes or magnitudes, a function specifies a way that, in some context, one measurement might depend upon another. If, as in many physical settings, that relationship be causal, then the function quantifies a potential causal relationship.

Particular functions are related to each other by addition and by multiplication by numbers. They are a *system* of measurements; they are a mathematical domain. But are they also a mathematical set? Can one speak meaningfully of a set of continuously differentiable functions? Can one specify and delimit the *range* of possible functions and the *respects from which two functions can differ* from one another?

Continuously differentiable functions are a subcategory of functions, in general, but this qualification, of being continuously differentiable, is irrelevant to the main issue. If I can show that functions, in general, can be taken as a set, my argument will also apply to continuously differentiable functions. This, for two

reasons: First, because the definitions for addition of two functions

and for multiplication of a function by a number, apply to functions

generally and, second, because, I have already qualified

continuously differentiable functions as, in effect, a sub-system of

functions taken generally.

Consider, first, that the *graph* of a realvalued function of

one variable, from the perspective of analytic geometry, is simply a

set of ordered pairs of numbers. A function is specified completely by its graph. A point set *qualifies* as the graph of a function of x provided that each x value occurs only once in the set of ordered pairs. If a function is defined over a set X (such as an interval) then the set of x-values in its graph consists precisely of any x *contained* in X ($x \in X$). In other words, the graph lies completely over the

interval and a vertical line through any point on the interval will

intersect the graph in exactly one point. Two functions f and g are

different if and only if their *graphs are different*; if and only if there is a point x somewhere on the interval such that $f(x) \neq g(x)$.

The range of possible functions is completely specified by

the criterion I have stated on its graph. Any meaningful

characterization of particular functions, within this range,

unambiguously isolates certain functions bearing determinant,

circumscribed, relationships to other functions in the mathematical

domain. The mathematical domain of functions is a set.

This characterization of functions presupposes the

mathematical domain of analytic geometry. It presupposes that one

can specifically identify, for mathematical purposes, (i.e., purposes

of measurement) the points in the plane with ordered pairs of numbers. But to view ordered pairs as a mathematical domain presupposes, in turn, that one has identified real numbers as a domain. The hierarchy starts with number, proceeds to ordered pairs of numbers, then to analytic geometry, and, finally to functions.

Now restrict attention to continuously differentiable functions. The difference between two such functions is completely specified by the set of points on which they differ and by the amounts of those differences. A characterization to distinguish certain continuously differentiable realvalued functions from the others specifies a set of continuously differentiable functions. For a mathematical domain to count as a set, one must specify a range of possibilities and specify the respects in which two measurements or points within that domain can differ. To unambiguously characterize certain measurements or points within a mathematical domain, as opposed to others, is to specify the set of measurements or points so characterized, as opposed to the complement set of measurements or points that are not.

Once again, the hierarchy starts with numbers and presupposes that numbers are a system of measurements, that numbers constitute a mathematical domain and, finally, that the

mathematical domain of numbers can be regarded as a set.
Mathematical domains are everywhere in mathematical

thought. In the sense I have given here, the mathematical

structures of 19th and 20th century mathematics, structures such as vector spaces, groups, matrix algebra, topological spaces, function

spaces, and differentiable manifolds, should be thought of as

mathematical domains. In reality, none of these “structures” was simply invented out of thin air. Each represents the conceptualization of a certain category of quantitative relationships

requiring special study, a study that was pursued, initially, because

it provided an approach to an existing mathematical problem or scientific investigation. Such structures arise either as a class of

existents regarded geometrically or, less directly, as a system of

measurements. In either case, there is a universe of discourse of the

kind that I have characterized as a mathematical domain.

Geometric Domains

One thinks of Cartesian coordinates

spatially, but, as

applied to the world, Cartesian coordinates pertain, more generally,

to arrays of possible values of a pair (or, even more generally, an ntuple) of quantities. I say quantities, in part, because to restrict the applicability of numerical coordinates to magnitudes would be a

little too narrow. As the simplest example, coordinates are routinely

used to represent position. But, although *relative* position along a single dimension is a magnitude, *position*, as such is not. One chooses a point of origin, a point of reference, and measures from

the point of reference. For numerical coordinates to be applicable

in the full sense, it suffices to say that the *difference* of two values of the x coordinate or of the y coordinate represents a magnitude, a

magnitude with either a positive sense or a negative sense.

In application, coordinates can represent possibilities or

configurations. For example, a magnitude, such as weight, a

quantity such as voltage potential, or a difference of two

magnitudes such as the difference between two volumes, can all be

measured by a point on a real line. In these examples, the real line

functions as a one-dimensional coordinate system. Every point on

the real line represents a possible status of the quantity under

consideration.

In general, the x and y coordinates of a coordinate system

may be used to represent different kinds of quantities. For example,

one could represent the weight and volume of an object by,

respectively, the x and y coordinates of a point in the Cartesian

plane. The configuration space of all possible combinations would

be represented by a particular subset of the plane, namely the first

quadrant, consisting of positive values for weight and positive

values for volume

values for volume.

One could also use the x coordinate to measure the radius

of a circle and the y coordinate to measure its area. Since these

attributes are related to each other, the configuration space of

possible circles is a curve consisting of points $(x, y) = (x, \pi x^2)$ where $x > 0$. This curve is expressed by the equation $y = \pi x^2$ for positive x.

In three dimensions, the possible positions of a particle are

represented by a three-dimensional coordinate system. The

configuration space of the particle is the entire three-dimensional space.

The configuration space of a three-dimensional

object is

somewhat more complicated. First, its center of mass could take

any position in three-dimensional space. Secondly, one must

account for the direction in which a chosen major axis is pointing.

Since any direction can be characterized by its intersection with a

sphere, the configuration space of the chosen axis can be measured

by a sphere (or by coordinates, such as latitude and longitude, used

to measure position on the sphere.) Finally, one can rotate the

object up to 360^0 around its major axis and the particular degrees of rotation can be measured (in radians) by points on a unit circle.

So to describe a possible configuration of a solid object requires

three coordinates to identify a point in space (R^3), two coordinates specifying a

point on a sphere (S^2) to measure the direction of the major axis, and one point on a circle (S^1) to determine the angular orientation around the major axis. One says that the configuration

space is $R^3 \times S^2 \times S^1$ to represent the respective domains of these three measurements.

As a somewhat similar example, consider the hour hand of

a clock. Its configuration space is a circle. Now suppose that a

second hand were attached to the end of the hour hand and could

rotate freely on the clock surface around that point. For any

particular position of the hour hand the second hand has a

configurations space, also of S^1. The entire apparatus has a configuration space of $S^1 \times S^1$. The first coordinate, from the first circular axis, represents the position of the hour hand and the

second coordinate represents the position of the second hand

attached to the end of the hour hand.

In general, any constellation of measurable attributes of a

physical object can be thought of occupying a position in a

configuration space. To regard a constellation of *measurements* pertaining to an object or a physical situation *geometrically* is precisely to consider its configuration space. Indeed, when a

physicist or an engineer chooses coordinates to measure these

various attributes, that's essentially what he is doing. And when one

applies physical laws to study its motion, as in the Lagrangian formulation of mechanics,[23] the trajectory can be thought of as a curve in the configuration space, one parameterized by time.

A geometric domain is a theater, a

space, where things

happen. A geometric domain can represent a configuration space,

as I just described, but it can also provide an occasion for other

related measurements. For example, the motion of a projectile in

three-dimensional space can be measured by a function from a time dimension into R^3. If a motion is confined, say, to a sphere (S^2), such as the surface of the earth, then S^2 is thought of as the geometric domain and the motion is a function from R (real number representing time) to S^2. A geometric domain can also represent a location for various shapes, as, for example, the plane in

Euclidean geometry or the set of coordinate pairs in analytic

geometry. And a shape, such as a circle, can represent the *path* of an object within its configuration space. In the case of analytic

geometry, such shapes are often characterized by equations, such as

$x^2 – y^2 = 9$ or by graphs of functions, such as $y = x^2 – 2x – 5$.

One may also consider functions defined on a geometric

domain taking values in another mathematical domain. One might,

for example, consider a function $T = g(x, y)$ to represent the

temperature at each point (x, y) in the Cartesian plane. Here, R^2 is the geometric domain and the function takes values in the domain

of real numbers.

Functions on a geometric domain need not be numerical.

For example, fluid motion in three dimensions might be

represented by a function on R^3 of the form V = v(x, y, z), where *V* represents the velocity (speed and direction) of the fluid at a point

(x, y, z) in R^3. Since a velocity vector is measured by three coordinates, this function has values in R^3.[24]

Finally, one measures various properties of geometric

domains: properties such as the surface area of a sphere or the

curvature of a sphere.

All of these kinds of relationships might apply to other

more complex geometric structures, curved surfaces, for example,

or even to more complex domains such as the configuration space

of possible positions of a solid object, which I found earlier to be

$R^3 \times S^2 \times S^1$.

In every case, a geometric domain provides an abstract

perspective on something measurable that is being investigated

with respect to the relevant measurements. As such, a geometric

domain always possesses some kind of structure, some kind of

measurable relationships between points in the space. One studies

it in all the ways I've mentioned, relating it to other domains. One

considers relationships among the points in the domain, but also

considers functions from other domains, such as the trajectory

example, or functions from the geometric domain to other

mathematical domains, as in the temperature and fluid velocity

examples.

A geometric domain is an abstract mathematical

representation of a category of constellations of measurements,

pertaining to some object or organization of objects, measurements that pertain to and specify the relevant, recognized differences relating to such objects.

A system of measurements can also be treated

geometrically, as a geometric domain. When one does so, as in the real number line, one, thereby, studies relationships among the

measurements in the system, relationships that they have by virtue

of the things, e.g., the nature of the magnitudes that they measure.

One can always take a geometric perspective on any mathematical

domain. But the converse is false; a geometric space, in general, is

not a system of measurements. But, in one way or another, it is

measurable.

Sets play two main roles in geometry. First, they can simply

be a way of marking off a particular area for some reason. Typically

a set serves as the domain of a function within a larger

mathematical domain. And, secondly, a set can arise as the set of

points within some mathematical domain where something

interesting happens. One may, for example, be interested in the set

of point for which a particular function is zero, or where it fails to

be defined, or is discontinuous or, conversely, is continuous. For

geometric domains in general, as in all cases, a set is a means of isolation.

Derived Mathematical Domains

Standard set theory provides a number of ways to create new sets by combining previously identified sets in various ways.[25] Most, perhaps all, of these "operations" are valid if applied in an appropriate context. In particular, these techniques can be used, and, in practice, generally are used, to derive new mathematical domains from previously identified mathematical domains. In at least the cases I discuss here, these operations, as applied to sets within the respective domains, yield new sets within the derived domain. So any mathematical domain derived in this way from domains that are also sets can, itself, be considered a set.

The burden of this section is to identify, explain, and justify the most important of these derivations and, by so doing, to indicate an approach to such explanations and justifications that can be applied more generally. My purpose here is to explicate the key operations; I do not weigh in on such controversies as the Axiom of Choice.

Throughout, I will assume, without notice, that every domain that I use to derive other domains can be regarded as a set.

In particular, I will take such domains to consist of elements, albeit elements that are related to each other in various ways. Secondly, any set theoretic operations that I will discuss with regard to mathematical domains apply to subsets of their respective domains, as well. And these derived sets will be subsets of their derived domains. In all of these cases, these derived sets will represent a distinct range of possibilities, within a mathematical domain, in which all elements are unambiguously distinguishable.

Obviously, a derived mathematical domain is hierarchically dependent on the domains from which it is derived, to the extent that there is no more direct way to characterize it.

Assume that A and B are mathematical domains. Then the product domain A × B consists of ordered pairs (a, b) where a ϵ A and b ϵ B. (Read: “a is an element of A and b is an element of B.”)[26] Any characterization of specific ordered pairs in A × B must involve a characterization of which elements *a* go with which elements *b*. Any ordered pair (a, b) so characterized consists of two “coordinates,” each of which can, independently, be unambiguously distinguished within their respective sets A and B. Therefore, each ordered pair (a, b) is unambiguously distinguished as a particular ordered pair, as an element of A × B. If one can unambiguously identify an element in the set A and a second element from a set B,

one, thereby, identifies an ordered pair consisting of the two elements.

What does this mean and why would one care?

The simplest example of this derivation is the Cartesian plane consisting of ordered pairs of real numbers. In general, the domain A represents a kind of measurement, say, in application, the weight of an object and B represents another measurement, say its volume. To consider A × B as a domain is simply to consider both measurements at once as separate attributes of an object. In the abstract, one considers A and B to represent, say, two systems of measurements, or constellations of measurements, that might measure separate attributes of an object or of a particular physical context.

In general, to take the product of two mathematical domains is, in essence, to consider a larger set of measurements, to measure along more dimensions. One measures both A and B.

When one adds dimensions by taking a Cartesian product one is not *constructing* something; one is adding a *perspective* from which to measure something.

To take a second example, the domain A might represent time, as measured by calendar days, and B might represent the closing average of the DOW. The set A represents all possible days within a particular range and the DOW represents all possible

within a particular range and the DOW represents all possible

closing averages. The product A × B is the configuration space of

possible combinations. The graph of actual closing averages, by

day, is a particular subset of the range of possibilities, a subset of

the configuration space specified by A × B.
In each of these examples, I speak of an application of a

mathematical abstraction to a particular concrete. Considered as a

mathematical domain, one does *not distinguish* between various applications of that abstraction to particular concretes. One simply

remembers that the *ultimate meaning* of the abstraction consists in these concrete applications, consists in the quantitative

relationships they are used to identify and embraces all the

concretes subsumed under the abstraction.
The concept of products of domains can be extended to

multiple products, such as A × B × C consisting of ordered triples

(a, b, c) belonging, respectively to A, B, and C. Everything I said about A × B applies to A × B × C, as well. Alternatively, one could

look at A × B × C as being defined recursively, by

A × B × C = (A × B) × C

Here, (A × B) × C is the product of the domain (A × B) with

the domain C.
One very important case is the case for which A = B = C

(etc) = R, where R is the mathematical domain of real numbers.

The n-fold product R × R × … × R (n times) is written R^n. An equation in n

unknowns or variables is an equation on R^n.
As another example, I earlier designated the configuration

space of possible spatial positions of a three-dimensional solid

object as $R^3 \times S^2 \times S^1$. In this expression R^3 is the 3-fold product of the real number line R. The total expression is the product of three

sets, R^3, S^2, and S^1. But one word of warning: S^2 is the twodimensional sphere (as in globe) and S^1 is the standard notation for a circle. It is *not* the case that $S^2 = S^1 \times S^1$. ($S^1 \times S^1$, as it turns out, is a torus, i.e., the shape of a donut.)
A second very important example of a derived domain is a

domain of mathematical functions. If *A* and *B* are mathematical domains, functions from A to B, of a specified type, taken as a

whole, can also be regarded as a domain. As a mathematical

abstraction, a function from A to B is characterized by some sort of

rule or specification, call it *f* for this discussion, such that to any element a ϵ A, corresponds a unique element f(a) = b ϵ B. One also

writes $f:A \rightarrow B$ and $f: a \rightarrow b$. As I discussed earlier, regarding real

valued functions on an interval, a function $f:A \rightarrow B$ has a graph

consisting of points in A × B of the form (a, f(a)), a ϵ A. Since f(a)

must be specified uniquely, the graph of the function f contains a

unique point in A × B corresponding to each a ϵ A.[27]
The trajectory of a point mass in three-dimensional space,

for example, is given by a function from R to R^3, where R represents time and R^3 represents three-dimensional space. The distribution of temperature in a solid object is given by a function from a subset

A of R^3 to R. The subset A represents the space occupied by the solid object and

R represents temperature. Again, these examples

are two of the concretes subsumed by the related mathematical

abstraction.
In that a function relates mathematical domains, it

constitutes a kind of measurement, as I outlined in my example of

continuous functions.
Since every function has a graph, a function is

characterized by a particular set of points in

A × B, namely {(a, f(a)), a ϵ A}. Since every subset of A × B is

distinguished from every other subset, every *set* of subsets of A × B

is distinguished from every other set of subsets of A × B.
Two subsets of A × B are different if and only if there is an

element (a, b) present in one of the subsets that is not present in

the other.
A set of functions relating a pair of domains A and B, thus,

corresponds to a set of *subsets* of A × B of a particular kind. A function can be thought of as given by a graph. A set of functions is

given by a set of graphs. And, finally, two *sets* Y and Z of subsets are different if there is a subset in Y that is not contained in Z or vice

versa. Graphs are simply a special case, a special kind of subset of A

× B.
Functions from a domain A to a domain B are a

mathematical domain. As a domain, they are identified by a domain

of graphs, by a particular subset of the set of all subsets of A × B.
If B is a system of measurements, then one can induce a

similar system of measurements on any domain of functions taking values in B. This is because any operation of elements of B is inherited by functions taking values in B. For example, if there is a way to add two elements in B (to get another element of B), then one can define the sum (f + g) of two functions from A to B by specifying its value at each point in A just as I did for realvalued functions. To wit, for any a ϵ A, the value of (f + g) is given by (f + g)(a) = f(a) + g(a).

My argument that graphs of functions constitute both a set and a domain applies without change to subsets of a domain. The range of potential subsets of a set is strictly circumscribed and they differ unambiguously from each other. So any meaningful characterization of certain subsets, as opposed to other subsets of the same domain, is a set. In general, the set of all subsets of a mathematical domain, the socalled *power set*, is both a mathematical domain and a set.[28] A subset of the power set is a set.

For example, the set of all lines through the origin in R^n is a mathematical domain known as projective space. Specifying a line is like identifying a direction – except that one does not distinguish between the two directions available on each line.

A second example is the set of all triangles in the coordinate plane. Considered as subsets of, say R^2, the position of a triangle is specified by the positions of its three vertices, without

regard to ordering of the vertices. Any three non-collinear points (points that don't all lie within a single line) determine a triangle located somewhere in R^2. Three coordinate pairs, six coordinates in all, subject to the non-collinearity restriction, determine a triangle in the Cartesian plane.

Subsets of mathematical domains constitute another category of derived mathematical domains. A subset of a mathematical domain is, obviously, a set. However, a subset of a mathematical domain may not be a system of measurements even if the parent domain is a system of measurements. For example, the set of odd numbers is not a system of measurements because an odd number plus an odd number is not an odd number. But the set of even numbers counts as a system of measurements because an even number plus an even number is an even number.

A good example of a geometric sub-domain is a sphere in three-dimensional space. The coordinates of the ambient space serve as one way to provide coordinates on the sphere. On the other hand, as a specialized study, one can study spheres without regard for the ambient space in which it sits. One can, of course, as I suggested, use the coordinates of the ambient space as a convenience, but this is simply a possible means of measuring the sphere that does not impact the inherent nature of the domain.

Another very common example is to restrict a

mathematical domain to a region such as a disc in R^2 (interior of a circle). One normally restricts to such domains, perhaps just

temporarily, as a methodological device, to study functions that are

intentionally defined only on the region. For example, one might do

this to find a local solution to a differential equation, based on some

kind of approximation technique, with the ultimate goal of

extending the solution to a larger domain. When one does so, one

considers the sub-domain as part of the larger domain. In this case,

the mathematical domain is also, in the standard terminology, the

domain of the function.

A final way to derive a domain from an existing domain is

to simply omit some of the measurements that are distinguished in

the starting domain. This means that one no longer makes all of the

distinctions that one had been making previously. For example, in

regards to my earlier example of triangles in the Cartesian

coordinate plane, one can ignore the orientation and placement of

each triangle. Two triangles are equivalent, from this perspective, if

they are congruent, if they can be matched up with corresponding

sides being equal. A triangle in this domain is characterized by just

three numbers, the lengths of its three edges, subject to the triangle

inequality and without regard to the order of the sides.

Standard treatments deal with this phenomenon of

measurement omission by a certain settheoretic construction

measurement omission by a certain settheoretic construction.

These treatments isolate a certain class of non-intersecting subsets

of the original domain. Then they assign each *element* of the original set to the *subset* consisting of all the other elements from which it is no longer distinguished. In the standard terminology,

this subset is called its *equivalence classes*. The new domain is characterized as the *set of such subsets*, i.e., the set of equivalence classes.
For further examples of this approach, see my discussions

of quotient vector spaces in Chapter 7 and of quotient groups in

Chapter 8.

Other Operations on Sets

There are other operations of sets that, contrary to official

set theory, only reasonably apply, within the context of a particular

mathematical domain, to sets within that domain. Most notably,

these operations are unions, intersections, and complements. As we

will see, however, within the ZF axioms my caveat does not arise

because the ZF axioms prescribe, in effect, the domain of all

mathematical sets, the entire domain recognizable as sets within

the scope of those axioms.

The union of two sets A and B consists of elements that are

either in A or in B. The intersection consists of elements that are in

both A and B. And, finally, the complement of B with respect to A

consists of points of A that are not elements of B.

Within a mathematical domain, such operations on sets are valid, logical, and straightforward.

On the OpenEndedness of the Set Concept

As I will pursue in the final section of this chapter, modern mathematicians, for a variety of reasons, conceived of set theory as a closed system in the early part of the 20th century. But the considerations that led to this historic development do not do justice to either why or how set theory came into being, why it applies in mathematics, or how new kinds of mathematical domains arise in mathematics. Settheoretic methods had arisen naturally in the work of Dedekind and even earlier[29] before the subsequent attempts of Cantor and others to institutionalize it as a purported foundation of mathematics. Set theory did not arise in a vacuum. It originated in a mathematical context, but ultimately took the form that it did because of a wider philosophical context. I will discuss that development at the end of this chapter, but, to my mind, the modern form that set theory has taken is an unhappy accident of history, driven by broadly accepted philosophical dispositions rather than mathematical needs.

In my view, the concept of a set, like concepts in general, is openended. Sets arise in specific mathematical contexts. From a

reality-based perspective, to establish something as a set requires the kind of considerations I have adduced in this chapter.

In practice, one will usually find ways to derive new sets from already established sets. Indeed, mathematicians have adopted the standard settheoretic axioms, based on a) their formal derivations of the real number system from those axioms and b) their belief that all sets of mathematical interest can be derived, from a purely formal perspective, from the real numbers via settheoretic operations provided within the axiom scheme.

But mathematics is not just a formal game. Sets arise in relation to problems in measurement, problems arising either in their applications or in mathematical problems involving indirect measurement. They arise because they serve a particular function. And it should not be a given that, to qualify a mathematical domain as a set one must “derive” it from the standard settheoretic axioms or any other set of settheoretic axioms.

In chapter 8, I discuss the development of finite groups. From the standpoint of formal mathematics, a finite group is a set because it can be identified with a set of integers, which are already recognized as a set. And, from a purely formal perspective, group multiplication is just a function that one associates with the set. But, from a conceptual, reality-based perspective, a finite group is

But, from a conceptual, reality-based perspective, a finite group is not a set of integers. Nor can one “identify” a set of group elements with a set of integers without stepping outside of one’s axiom scheme. For there is nothing anywhere within the standard axioms of set theory to relate anything within its domain to any external existent or phenomenon whatever. One-to-one correspondences, within the compass of standard set theory apply only to objects that have already been identified as sets.

Set Theory and Point Set Topology

I return to the question that I asked earlier. What does it mean for one function to approximate another function? My answer will involve an interesting and non-routine application of set theory.

Set theory was unknown in the classical period and only implicit in analytic geometry. The great achievements of Newton, Leibnitz, Euler, Gauss, Cauchy, Riemann, and countless others, all predate the axiomatic set theoretic scheme of the twentieth century and even the naïve set theory advanced by such mathematicians as Dedekind and Cantor. The logical analyses that mathematicians relied on for millennia, including sophisticated concepts, such as Cauchy’s definitions of

limit and

continuity, all applied the

Aristotelian logic dating back to ancient Greece. The wonderful

mathematical achievements through the end of the nineteenth

century were all achieved without any need for set theory or its

modern logical superstructure, as a formal discipline.

Yet set theory is more than just a convenience. Like every

mathematical abstraction, set theory provides a building block for

further discoveries, for new conceptual formulations and

identifications. And one of its major triumphs is point set topology.

Point set topology is *not* about sets; it's about a certain kind of *measurement*. But point set topology requires a sophisticated

application of set theory. And that application is left intact within

my revisionist view of a proper formulation of set theory.

I present this conceptual introduction to point set topology

for several reasons. First, topologies are important. Point set

topology provides the ideal framework to deal with the kind of

question that I asked earlier. Namely, what does it mean for one

function to approximate another function?

Secondly, granting the importance of point set topology,

my discussion will exhibit point set topology as an important, nontrivial application of set theory. It is one thing to validate a concept; it is another to show decisively why that concept matters, why it is

show decisively why that concept matters, why it is

important.

Finally, point set topology is one of those areas in twentieth

century mathematics in which an initial skepticism by a noninitiate is not out of place. It is reasonable to ask: How could such a theory as topology capture anything important about the world?

Why isn't point set topology a pointless floating abstraction?

Assuming that the concept is meaningful at all, why should

anybody care? These questions, in my view, have good answers,

but, unfortunately, it would be unusual to hear either these

questions or their answers from a contemporary mathematician.

The goal of this section is to show why point set topology is

important, where it comes from, and, from a this-worldly

perspective, what it actually means. And, in showing that, to show

why set theory deserves rehabilitation from a realist, this worldly

perspective, why a systematic framework of set theory is valuable

and meaningful.

As mathematicians look at it, topology is the general study

of continuity. This is certainly true, especially of more sophisticated

sub-specialties such as algebraic topology. And the concept of a

continuous function finds the most general and fundamental

expression in topology. But, in its initial motivation, at least from a

reality-based perspective, I believe the essential purpose of *point set topology* is to provide a general foundation for the study of

approximation in mathematics: How, in the most general context, should mathematics measure proximity? And that will be the theme

of my introduction to point set topology.

In general, to say that something approximates another

thing is to say that they are close in some way. To say that one

number approximates another number is to say that their

difference is small. To say that one point in R^3 is close to another point in R^3 is to say that the *distance* between them is small. If x is the point with coordinates (x_1, x_2, x_3) and y is the point with coordinates (y_1, y_2, y_3), then, according to the standard distance formula (based on the Pythagorean theorem, as applied to three

dimensions), the distance D(x , y) between x and y is

$$D(x , y) = \sqrt{((x_1 - y_1)^2 + (x_2 – y_2)^2 + (x_3 – y_3)^2)}$$

How does this approach to approximation work out for

numbers and points? And how can it be applied to functions?
Assume a sequence of numbers $a_1, a_2, a_3, \dots , a_n, \dots$. One says that the sequence converges to a number a if, for any standard

of precision $\varepsilon > 0$, there exists a number N such that $n > N$ implies

that $|a_n – a| < \varepsilon$.
In other words, any required level of precision can be

achieved by looking at a finite number of terms in the sequence.

This is, indeed, part of the point of the approach to irrational

numbers presented in Chapter 4.
Convergence works pretty much the same for points in

space (R^3). To apply the distance function D(x,y) just defined, a sequence x_1, x_2, x_3, … converges to a point X if and only if for any standard of precision $\varepsilon > 0$, there exists a number N suchthat n >N

implies that $D(x_n, X) < \varepsilon$. Again any required level of precision can be achieved by looking at a finite number of terms in the sequence.
This criterion can be reformulated in a useful way. And that

reformulation is useful because it provides a bridge to a more

general treatment of proximity, namely, the treatment formalized

in point set topology.
First, some definitions: By an open ball of radiusε around a

point X, one means the set of B_ε consisting of points with a distance *less*thanε from X.Tosay thatthe ballis *open* means that one does *not* include the surface of the ball in the set B_ε.
These terms having been established, one can reformulate

the convergence criterion as: For any open ball around X, there

exists a number N such that every term in the sequence after the

Nth term is inside the ball.
In applying this criterion to obtain a suitable

approximation within a particular context, one chooses a ball so

small that there are no material distinctions between any of its

points. That is, each open ball embodies a *potential* standard of precision.
But a particular open ball is a standard that applies only to

the points that it contains. It says and implies nothing about

precision for points outside the ball. This use of balls is a way of

localizing the global precision criterion specified by $\varepsilon > 0$. A ball of radius ε

arounda pointXimposes thesameprecisionstandardasa

similar ball of the same size around a point Y. But that standard is

applied independently at each point.
To make the discussion more general, one can even say

something that may seem weaker at first glance: for any set U

containing an open ball around X there exists a number N such that every term after the Nth term is inside the set U. I say “may

seem weaker” because making sure a point is inside the ball is enough to guarantee that the point is inside the larger set U that

contains the ball. And, conversely, one could, in particular, choose

U = B so the statement about U includes my earlier statement about

B.
Notice something special about an open ball: for any point

y inside an open ball B there is a smaller open ball B_y centered around the point y and contained in B. One writes this

$$y \in B_y \subset B$$

This property holds for

any union of open balls and *any*

finite intersection of them. The case of unions being more obvious,

I elaborate the case of a finite intersection. Consider a point Y

contained in two different open balls, B_1 and B_2. There is a smaller ball around Y contained in the first ball, B_1, and there is also a ball around Y contained in the second ball, B_2. The smaller of these two smaller balls is contained in both of the larger balls, B_1 and B_2.

A set

U with this property, that every point in the set U is

contained in a ball, centered at that point, that is *also* contained in U, is called an *open set*. And the generalization I just enunciated for open balls holds, more generally, for open sets. Any

finite intersection of open sets is an open set and *any* union of open sets is an open set.

Any point in an open set contains an open ball around it

that is completely contained within the open set. Taking this as the

defining condition, any subset of R^3 is either an open set or it is not.

Both R^3 and, by convention, Ø (the empty set) are taken to be open sets. (The defining condition for an open set may be considered to

hold vacuously for Ø.) So the set of all subsets of R^3, perhaps unexciting in itself, has a very important subset, namely the set of

open sets that are subsets of R^3.

As my remark above indicates, the standard of convergence

enunciated above relative to open balls would hold just as well,

would be completely equivalent, if I were to say: “For any open set

U around X, there exists a number N such that every term after the

Nth term is inside the specified open set U.”

An open set, as it relates to the points it contains, is a

standard of precision. Taken as a standard, invoking an open set U

is to say that there is no material difference among its points. Open

sets are like a system of open balls. They have the same

mathematical purpose, but are more generally applicable to other

contexts.

To translate a criterion of

convergence in the way that I

have illustrated, so as to utilize a *system* of open sets, is to relate that criterion to a broader conceptual framework. This broader

conceptual framework, as it turns out, applies to every convergence

criterion that arises in mathematics.

A specification of the open sets of a geometric domain is a

determination of the meaning of proximity, a characterization of

what it means for one point to be close to another point. It is a

standard of relevance, of the respects in which differences matter.

Open sets are a system of filters. To choose an open set U is to

choose a standard of proximity in regard to the points that it

contains. The open sets, in their totality, provide a system of such

standards. A specification of the open sets of a geometric domain is called a *topology*.[30]

Again, I bring up point set topology for three reasons. First,

it is an interesting, important, and nontrivial use of set theory,

properly understood. Secondly, it *is* important. And thirdly, it is not at all obvious to most people why it *should* be important. Why isn't the notion of topology unnecessarily abstract?

And, in this connection, I return to my first question: What

does it mean for one function to approximate another function?
The short answer is: approximation of functions can, does,

and should mean a lot of different things depending on the context.

There are many different *respects* for functions to be relevantly close to each other and all of these respects serve particular needs.
And this leads to part of the answer to my question on

topology. Some of the ways of measuring the proximity of two

functions do not involve assignment of a *number* to measure

proximity of the two functions. Numbers may still be involved, but

one cannot always use a *single number* to capture and delimit all of the important, relevant ways in which two functions can differ. Yet

there are useful, more general ways to measure proximity. And

every one of those ways implies a particular kind of topology, a

particular determination of the open-set-measurements of

precision requirements, a determination that reflects the *kinds of differences* that are relevant to a certain kind of pursuit and

provides a system of filters sufficient to specify any required level of

precision.
Obviously, some examples would be helpful. I start with the

simplest: pointwise convergence of a sequence of realvalued

functions. In this kind of convergence, one says that a sequence of

realvalued functions f_1, f_2, f_3, … ,f_n, … converges to a function f on some domain if and only if, for every x in the domain and for every

$\varepsilon > 0$, there exists an N such that $n > N$ implies that $|f_n(x) - f(x)| < \varepsilon$.
Notice that, for pointwise convergence, the value of N

required to guarantee a specified level of precision *depends on one's choice of x*.[31] Suppose, for example, that f_n is the function $f_n(x) = x^n$ on the *closed* interval from 0 to 1.[32] (The adjective, *closed*, signifies that one includes the two endpoints of the interval within the

domain of the function.) Every function f_n in the sequence is a continuous function and, indeed, an infinitely differentiable

function. But, although the sequence has a limit, this limit function

is not continuous. For every $x < 1$, the limit is 0; for $x = 1$, the limit

is 1. The limit takes on no other values besides 0 and 1. This lack of

continuity of the limit function reflects the fact that, the closer one

gets to $x = 1$, the slower the sequence $\{x^n\}$ converges to zero.
Clearly, if, as is often the case, one wants a sequence of

continuous functions to converge to a continuous function, a

stronger notion of convergence is needed.
This pointwise-convergence approach has another,

related, *price*. If all one cares about is the value of the limit at a specified finite, isolated set of points $\{x_i$ for i between 1 and m$\}$, then pointwise convergence will provide a value of N that will

simultaneously satisfy one's standard of precision $\varepsilon > 0$ for each of those points. Simply find a suitable number N_i for each point x_i and then take the largest of these. Let N = maximum of $\{N_i\}$. Then, for all $n > N$ and each x_i, $|f_n(x_i) - f(x_i)| < \varepsilon$.
One can achieve the required precision at a pre-specified

handful of points. But, for pointwise convergence, that's all that

one can hope to achieve in a finite number of steps.

one can hope to achieve in a finite number of steps.
Before looking for alternatives, it is helpful to look at

pointwise convergence topologically: What are the open sets

corresponding to this notion of convergence? In this case, there is

no concept of a distance between two functions; there are only

distances between the values of these functions at particular points.

Numbers are involved, but each number applies to only one point

in the domain of the functions. But, despite the lack of a concept of

distance, I can use the same general approach to defining the

topology that I used in the simpler cases that do have a concept of

distance. I simply start with my convergence criterion, just

enunciated for a finite set of points.
Accordingly, choose a finite set of points $x_1, \ldots, x_m$. For each point x_i, choose a number y_i. And choose $\varepsilon > 0$. Think of y_i as a *potential* value of a limit function f at the point x_i. Consider the set of functions g such that $|g(x_i) - y_i| < \varepsilon$ for all values of i. Then I declare the set of such functions g to be an open set, a set of

functions that satisfy the standard of precision set by the value of ε

with respect to the chosen potential values at the selected points.
This open set is somewhat less than what one might

consider a standard of precision; it is a *local* standard of precision.

It is a standard of precision

relative to the values y_i at

corresponding points x_i. The precision *standard*, the local measure of proximity is specified by $\varepsilon > 0$. The x's and the y's pertain, first,

to *where* that standard is applied (the x values) and, second, *what the function is close to in those places* (the y values). Taken as a

whole, this set of conditions is what mathematicians call a *filter*.

One should think of invoking the filter like this: To satisfy it, means

that, as far as the values of the function at the x-values, $x_1, \ldots, x_m$, are concerned, the values of the function do not differ materially

from the corresponding y-values y_i. If f were a different function with nearby values at the prescribed x values, then, by this

precision standard, g would be indistinguishable from f,

indistinguishable because they are both contained within the same

filter that sets the standard. The set of functions satisfying the

condition are all indistinguishable from each other as far as this

particular filter is concerned.

But one filter is not enough to evaluate convergence of a

sequence. Pointwise convergence means converging at every point.

One needs a *system* of such filters, covering every point in the domain and covering all potentially relevant degrees of precision.

These filters play the same role as the open balls around

points in R^3. In the case of R^3, one thinks of a series of balls, of varying radii, around every point in R^3. These open balls, as I have already remarked, are a system of filters.

Thus, to continue with pointwise convergence, there is a

similar open set, a filter, for every choice of a finite set of x-points,

of corresponding y-points, and of a positive value of ε. Taken

together these sets comprise a *system of filters*.

A system of filters provides the form in which one expresses a standard of precision. The filters specify the kinds of differences between *functions* that one considers to be relevant. Taken together, these filters represent every available precision standard and every possible application of these precision standards within the overall umbrella of pointwise convergence.

To define a topology means to specify all of the open sets. But that's actually straightforward once the system of filters has been established. The real work of defining a topology consists in identifying a system of filters: Any union of filters is an open set and any finite intersection of open sets, including filters, is an open set. More definitively, to paraphrase an earlier statement involving open balls: An open set is characterized by the fact that any point (i.e., particular function) in the open set is contained in a filter that is, itself, completely contained within the open set.

Keep in mind that, in this paragraph and, in general, relative to elements of open sets and filters, "point" refers to a function, since it is a *function* that we are, in this context trying to approximate.

This kind of discussion can get confusing because one approximates a function f by reference to the values $y = f(x)$ that the function f assumes at various points x in the domain of f. So the word point refers, depending on the context, either to a function f

or to a point x in the domain of a function f.
This is a standard way and the usual way to define *a* topology. However, this *particular* topology, as I've indicated, is not a great way to measure convergence of functions. But an

apparently small change in the citerion of convergence results in a

much stronger and, where feasible, a much more useful kind of

convergence, namely, socalled *uniform convergence*.[33] It goes like this: One says that a sequence of realvalued functions f_1, f_2, f_3, …

,f_n, … converges to a function f, defined on some domain, if and only if for every $\varepsilon > 0$, there exists an N such that $n > N$ implies that

$|f_n(x) - f(x)| < \varepsilon$ *for every x in the domain*. Notice that, for uniform convergence, N does *not* depend on one's choice of x.
Obviously, uniform convergence implies pointwise

convergence. If a sequence of functions converges uniformly, then

it necessarily converges at each point.
For a sequence satisfying this criterion, if one's standard of

precision is $\varepsilon > 0$, one can achieve *this* precision at one stroke, simultaneously for all values of x, by finding a large enough value of

N.
Related to this topology, one also has a kind of measure of

distance between functions. Namely, one expresses this measure by

the expression

$$D(f - g) = \sup\{|g(x) - f(x)|\}$$

where the settheoretic function sup is applied across all x

in the domain of f and g. In this expression, *sup* is a technical term for specifying the largest number in a set of numbers. But,

technically speaking, it really specifies the smallest number that is either greater than or equal to every number in the set. So, for example,

$\sup\{x < 2\} = 2$

This says that 2 is the smallest number greater than or equal to all of the numbers that are strictly less than 2. The need for this formulation consists in the fact that 2 is not part of that set of numbers.

It will frequently turn out that the value of this “distance” is infinite for certain pairs of functions. Consider, for example, the two functions $f(x) = 2$, and $g(x) = x$. For large values of x, the difference between these functions grows without limit.

Strictly speaking, then, mathematicians would not recognize this as a “distance function”. However, my interest centers on small distances between functions. The fact that differences between certain functions may be unbounded has no bearing on that pursuit. But it is certainly a complication, and a reason, even in this case, to take a topological perspective.

Unlike the case of pointwise convergence, a uniformly convergent

sequence of continuous functions, as it happens,

converges to a continuous function.

The topology of uniform convergence is very similar to the

topology for R^3 that I defined above. The filters are, as follows: Choose any function $f_0$and any $\varepsilon > 0$. Consider the set of all functions g such that

$|g(x) - f_0(x)| < \varepsilon$ for all x in the domain.

The set of all such functions g is taken to be an open set,

indeed, a filter. To specify a function $f_0$and an $\varepsilon > 0$ is to specify a particular filter. Any set that is derivable from finite intersections

and from unions, starting from the set of such filters, is an open set.

Convergence of a sequence for funtions f_n to f says that, for any open set (or filter) containing f, there is an integer N the open set

(or filter) contains all functions $f_n$for which n > N. For this particular system of filters, this is to say that f_n converges uniformly to the function f

There is also an interesting and very important

compromise between pointwise convergence and uniform

convergence, namely, uniform convergence on finite closed line segments.[34] The compromise is important because full uniform convergence is not always feasibly achievable, yet something better

than pointwise convergence is needed.

This version of convergence is analogous to pointwise

convergence in that a suitable value of N required to achieve a

standard of precision *does* depend on x. However, it's also similar to uniform convergence. To wit, one says that a sequence of

functions converges uniformly on finite closed intervals if and only

if, *for any finite closed interval*, the sequence of functions, restricted to that interval, will converge uniformly.

This is another topology without a distance function between functions. But it is much more useful than pointwise convergence. To the extent that one's real interest lies in a finite region, in the convergence over a finite interval, then, over this interval, one can achieve one's standard of precision, at one stroke, with a suitable value of N. One can achieve it in the same way, and for the same reason, that one can achieve it with uniform convergence.

These three examples are just a sampling of the range of important ways in which a sequence of functions can converge to a limit. There are many elaborations in the same spirit and they all relate to different important respects in which one function can approximate another. For example, when solving differential equations, one, generally, needs the derivatives of the functions in the sequence to converge to the derivatives of the limit. This requires putting bounds on the derivatives, as well, resulting in a more involved system of filters and a correspondingly more complex topology.

Finally there is an entirely different category of

convergence conceptions for functions. For this kind of

convergence, one doesn't really care about convergence at every

point, just at *most* points. One is much more interested in the *area* between two functions than in their differences at individual points.

As a classic example, a very general class of periodic

functions can be represented as limits of infinite series (of *Fourier series*) consisting of sine and cosine functions. This fact is the

mathematical basis for the overtone series in music.

In finding such infinite series, one's interest centers on

getting as close as possible to a particular periodic function, even

when that function doesn't happen to be continuous. The

appropriate standard of proximity varies accordingly. One's interest

here centers more on a "harmonic analysis" of a function as

appropriately approximated by a finite series of "overtones" – of

sine and cosine functions.

As a simple example, the function

f that is equal to 1

between 0 and 1, equal to -1 between 1 and 2, and periodic beyond

the interval between 0 and 2, can be represented as a limit of a

Fourier series. But, regardless of the value assigned to the function *f* when x is 0, 1, or 2, the *limit* function will be always be zero at all of these points, and so will the finite approximations to it.

In the study of differential equations, one cares a great deal

about convergence at individual points. In contrast, exceptional behavior of the limit at particular isolated points is largely irrelevant to anyone interested in discovering the Fourier series that approximates a periodic function. What matters much more is the difference of the area between the periodic function and its approximations. Or, more precisely, what matters is the average over the period, such as the interval between 0 and 2, of the square of the difference between the function being analyzed and its successive approximations by a Fourier series. In general, to define a *type* of convergence, to define a *topology*, is to state a *criterion of relevance and the system of filters sufficient to apply it*.

In short, point set topology is generally applicable to any measurement of proximity that may be required to study limiting processes. And in defining a common vocabulary to capture the essence of all convergent processes, point set topology has been able to find very broad principles applicable to all limiting processes, no matter what measure of proximity might be needed in any particular limiting process.

But convergence isn't just about finding limits; convergence is about finding good approximations. Indeed, from a reality-based perspective, finding limits *is* about finding good approximations. It is about finding approximations that do not

differ materially from whatever one is approximating.

One wants a good approximation in every respect that matters in a particular case. Whatever interval one selects as important, one needs, for a second degree differential equation, the value of the function to be close, but one also needs the first two derivatives to be close. If close means within 0.1%, then all of these values must be within 0.1% of the limit value. Within the required level of precision, and in the required respects specified by the topology, a qualifying approximation

is the solution to the

equation. The approximation is indistinguishable from the limit in any way that matters.

It is exactly this sort of relevance criterion that a choice of topology provides. And it does so in a general fashion that applies geometric concepts, i.e., of proximity, of convergence, and of continuity and connectivity, to any kind of measurement in mathematics for which approximations are needed and possible, for which proximity is meaningful, and for which it is meaningful for something to vary in a continuous fashion. Topology provides a consistent comprehensive way to capture any criterion of approximation, any criterion of relevance, and to apply that

criterion to specify a degree of required precision. Topology provides for a wide generalization of key geometric concepts such as continuity, connectivity, and limits. The applicability of topological concepts and methods is at least as extensive as the variety of continuous quantity. Wherever continuous change is possible there is continuous quantity and, accordingly, there is an application of topology. Point set topology is rightly regarded as one of the cornerstones of the study of mathematical functions that developed out of the differential and integral calculus.

The Modern Theory of Sets

If set theory has a defining characteristic, it is that a set has a precisely circumscribed membership and is completely determined by that membership. However that membership may happen to be described has no bearing. To deny this characterization is to speak about something other than sets.
In my own revisionist view, and within the context of a mathematical domain, I have insisted, as well, on this characterization and I have taken the function of isolation with the focus on the totality of elements isolated, within an appropriate context, as the essence of what sets are for and what they accomplish. In standard set theory, this viewpoint is captured in the

accomplish. In standard set theory, this viewpoint is captured in the socalled axiom of extension:

"Two sets are equal if and only if they have the same elements."[35]

Also, again within an appropriate mathematical context, I have explicated and offered rationales for standard settheoretic operations from which new sets may be constructed from old sets. And these constructions are the essence of what the modern theory actually provides. As Halmos puts it:

"All the basic principles of set theory, except only the axiom of extension, are designed to make new sets out of old ones."[36]

But, on the level of fundamentals, this is where the similarity of my perspective with the modern perspective ends.

Origins of Set Theory

The conception and use of sets is already implicit in the work of Reimann and others and explicit in the work of Dedekind.[37] But more than anyone else, Cantor was the father of set theory.[38]As Cantor defined it:

"By a 'set' we understand any collection M into a whole of certain welldistinguished objects … of our intuition or of our thought" (1895)[39]

Epple characterizes this as Cantor's "final definition of

sets."[40]
Implicit in Cantor's treatment and in subsequent

treatments is the view that there is a *universe of sets* that is somehow given once and for all. That viewpoint survives, in some

form, to this day and is presupposed by various axiomatizations of set theory.[41] The problematical notion of an alleged *set of all sets*, presupposes such a universe.[42]
Beyond his requirement that elements be "welldistinguished" Cantor recognized no restrictions on forming sets and, in practice, he contented himself with offering informal

descriptions of specific sets.[43] But the viability of this approach was soon beset by a number of paradoxes, some of them developing

directly out of Cantor's own work. The most famous and least

technical of these is Russell's paradox, a version of the liar's

paradox and generated by the question whether or not the set of all

"sets that are not members of themselves" contains itself as a

member. Upon reflection, one discovers that each possible answer

implies the other, contrary, answer.
In his classic *Axiomatic Set Theory*, Suppes reviews a

number of similar paradoxes arising out of the early naïve accounts

of set theory. Some of these paradoxes, he points out, "arise from

purely mathematical constructions," whereas certain others,

generally relating to the well-known liar's paradox, have a broader

semantic origin.[44]
Most striking, in these paradoxes, are the broad appeals to

all sets of a particular type, e.g., the "set of all ordinals."

all sets of a particular type, e.g., the "set of all ordinals."
A set is the product of a conceptual faculty. But arguments

invoking, outside of any prescribed domain, a "set of all…"

presuppose a kind of Platonic realism in which all sets have a kind

of intrinsic existence, prior to anyone's conceptions.
Notice that, from a proper referential perspective in which

assertions, even of possibilities, require evidence, these paradoxes,

dependent as they are on an intrinsic view of ideas, do not arise.[45] They do not arise because there is no Platonic universe of intrinsic

ideas. And reality does not contain contradictions. When one

appears to arrive at a contradiction, one resolves it by looking at

reality and checking one's premises.
Cantor's program was to ground all of mathematics on set

theory,[46] a program that was continued by his successors. But, if one wanted to avoid renouncing infinite sets altogether, some kind

of restriction on one's ability to call something a set was required.
In the standard diagnosis, there were two basic problems

with the early formulations of set theory. The first was the socalled

axiom of abstraction[47] that, in essence, said that any well-defined abstract

criterion specified a set consisting of the concretes

satisfying that criterion. A version of this axiom survives today, but

the surviving version only provides a way to define subsets of an

already-acknowledged set. In this form, it is known as the axiom of specification.[48] In rough terms, what distinguishes the surviving version is that the parent, the already acknowledged set, acts as a

genus.
The second problem is thought to have been a language

that was too rich. As Suppes puts it, “we avoid these paradoxes by severely restricting the richness of our language.”[49]
On such a diagnosis, one looked for ways to restrict the

ways that sets could be defined and to restrict the things that could

be said about them.
Ultimately, the standard, generally accepted, answer was

provided by offering (or prescribing) a list of axioms of set theory.

The most successful of these lists was published by Zermelo in 1908

and, in a slightly modified form (Zermelo-Fraenkel), is the standard

set of axioms (the *ZF axioms*) in use today by working

mathematicians.[50]
These axioms were designed to accomplish three things.

First, they needed, as much as possible, to accommodate the

standard kinds of arguments that were then commonly accepted

among mathematicians. Secondly, they needed to delimit the

universe of sets in such a way as to avoid the paradoxes that had

plagued earlier, informal, attempts at set theory. And thirdly, they

all needed to have enough intuitive appeal and wide enough

applicability to inspire assent from the mathematical community.
The resulting set of axioms asserts the unique existence of

the set having no elements, the socalled empty set, symbolized by

the symbol Ø. And the socalled *Axiom of Infinity* asserts the existence of a rather

peculiar set “which has Ø as an element and

which is such that if *a* is an element of it then the [union of *a* and {*a*}, namely {*a*,{*a*}},] is also an element of it.”[51] Here, the inclusion of {} around something reads something like “the set consisting of”

So the Axiom of Infinity is saying that if *a* is an element of this infinite set, then the set consisting of two elements, namely (the

element a and the *set* {*a*} whose only element is *a*) is also an element of the asserted infinite set.
Now these are two very remarkable sets. The first, namely

Ø, is essentially a mathematical convenience. And the second is an

infinity stew that mixes elements of sets with sets, with sets of sets,

and so on infinitum.
Aside from the Axiom of extension and these two posited

sets, every other settheoretic axiom prescribes some specific way of

generating new sets from existing sets.
In this, one basic underpinning of Cantor’s work has not

changed: “In presenting the ZF axioms it is presumed that the

domain of entities of which they are true is a universe of sets”[52]
But here is the paradox. The intent of the set of axioms is to

limit that universe to sets that can be derived from that set of

axioms. But that entire edifice includes but one element from which

to build, namely the set that has no members, the empty set. As

Tiles acknowledges, “it is a wholly abstract universe generated, as it

were, out of nothing.”[53]
A theory that once started as something completely openended secures its

foundations, and avoids contradictions, by

constructing its own universe from the set that has no members.

And, in a literal sense, the ZF set of axioms limits the world of sets

to sets that derive in some way from this base.
In my attempt to rehabilitate sets, I have emphasized

hierarchy, a hierarchy that starts with reality. The ZF axioms

provide an alternative hierarchy, as a way (it has been hoped) to

avoid contradiction, a hierarchy that builds from the empty set and

from an assertion of a certain related set provided by the Axiom of

Infinity. To directly assert the existence of *any* other set not derived from this base is to ignore the entire purpose of the axioms. It is to

risk the earlier paradoxes or, perhaps to encounter new ones. In

practice, then, within the *effective* scope of these axioms, *there are no sets except those that can be constructed, in the prescribed*

ways, from the empty set.
In the ZF framework, one does not have the foundation of a

referential theory to provide a defense against contradiction or to

provide a reference point to resolve any contradictions one might

find in one's thinking. So the ZF framework must carefully

prescribe a universe of its own. And one must be careful to never

venture beyond its walls.
How can one possibly use such a theory to provide a

foundation for mathematics?
And the short answer is that one utilizes the axioms to

construct a *formal* equivalent to known mathematics. And, as it turns out, from a formal standpoint, if one can somehow construct

an image of the natural numbers, one can *formally* generate, from that base, and with the aid of the ZF axioms, everything else.
So how does one construct the natural numbers? Like

this:[54]

$0 = \emptyset$
$1 = 0^+$ defined as $\{0\} = \{\emptyset\}$
$2 = 1^+ = \{0,1\} = \{\emptyset, \{\emptyset\}\}$
$3 = 2^+ = \{0, 1, 2\} = \{\emptyset,\{\emptyset\},\{\emptyset, \{\emptyset\}\}\}$

…

And so on, for all successors of 0.[55]
By the Axiom of Infinity, one can parlay this beginning into

an infinite set, one that can at least be *thought* of as the set consisting of all the natural numbers.[56] Then one constructs the socalled *Power Set* of that infinite set, consisting of all subsets of the infinite set.[57] And this, in turn, is enough to utilize the standard constructions of the real numbers, e.g., those of Dedekind and

Cantor that I discussed in Chapter 4. And, finally, the foundation of

the real numbers, one holds, is enough to construct everything else

in mathematics. When philosophers of mathematics say that the

status of the real numbers is the only interesting problem in the

philosophy of mathematics, this may be what they have in mind.
This entire enterprise has no existential referent

whatsoever, but, from a formal perspective, it, at least, purports to

recreate the entire formal structure of mathematics. I say

"purports" because such a formal equivalence has no meaning

within the province of the ZF axioms, which cannot refer, as such,

to anything, not even the natural numbers.
In any event, since the concern is to avoid contradiction,

and because, within a purely formal system, any contradiction that

arises would have to be a *formal* contradiction, mathematicians take away the following: Go ahead and think of numbers any way

you want to, including the way that you always have. Go ahead with

the constructions you've always done anyway. As long as something

can be constructed out of the real numbers, in the prescribed ways,

you can, formally, continue to live within the, hopefully safe, world

of Zermelo-Fraenkel set theory.
From the ranks of mathematicians, the only remaining

argument revolves around issues involving the validity of some of

these constructions, with the "intuitionalists" and "constructivists"

wanting, for various philosophical reasons, to place greater

restrictions on permissible set theoretic operations and on mathematical arguments than everybody else.[58] If the ontological status of mathematics remained unsettled, mathematicians came to

terms with set theory as an enterprise showing that they could

continue to go on doing what they had always been doing without

fearing the next settheoretic paradox. As Epple puts it:

"… it must be stressed that the philosophical issues

which were thrown up by the end of the crisis of

which were thrown up by the end of the science of quantity cannot be regarded as having been solved. If there is any consensus about the foundations of analysis, then this consensus consists in the pragmatic agreement that analysis should be practiced on the basis of the ontologically neutral axioms of set theory."[59]

Set theory, one should realize, is hardly ontologically neutral. The only sense in which set theory might be called ontologically neutral is that it ignores, or treats as irrelevant, any connection of mathematics to the world.

But, in practice, this attitude means taking numbers for granted and limiting one's settheoretic constructions to those that are on the list or that are derivable from the list prescribed by the ZF axioms.

General Comments

From a reality-based perspective this is simply crazy. It completely cuts off mathematics from any official relationship to the world. It ignores the context of how sets actually arise in mathematics, why they are needed, what they mean, and how is it that they actually provide the distinctions that they are designed to

provide.

To adopt the ZF axioms as a foundation of mathematics is

to abandon, on principle, any substantive content of mathematics.
At first glance, as it originated, set theory took the entire

universe, or perhaps the entire mental universe, as being

partitioned neatly into an array of interlocking sets. There is, in that

view, one universe of sets. For set theory maintains that any two

sets in the world can meaningfully be combined, by taking their

union, into a larger set, that one can, for example, combine

numbers, staplers, and emotional states into one infinite set.
Yet, to avoid contradiction, the ZF axioms offer a far more

constricted universe, one built entirely on the empty set and on

mental gyrations of things like: the set consisting of the set

consisting of the set consisting of … . Within that universe one

combines, freely, objects including different numbers of iterations

of this "set of" device. One purchases such freedom by basing the

entire universe of sets on, quite literally, nothing.
Official set theory replaces a hierarchy derived from

observation of the world with a different hierarchy, one based on

iterations of … nothing. It is a theory uniquely designed to support,

in a phrase contributed by Mary Tiles (after a statement of Hilbert's) Cantor's paradise.[60]

Why Mathematics has survived

In my view mathematicians, despite the absurdities of set theory, have continued, to this day, to do mathematics. But how is this possible? Assuming that I'm right, why has mathematics survived?

First, progress in mathematics has always been driven by the problems that mathematicians attempt to solve. Beyond the self-inflicted problems that arise in the philosophy and foundations of mathematics, these problems all arose, in some form, directly or indirectly, as problems of measurement. The major branches of mathematics of the $20^{th}$and 21^{st} centuries all have their roots in the 19^{th}, which have their roots in the 18^{th}.

Secondly, the entire machinery of set theory is taken, in a sense, with a grain of salt. Mathematicians treat numbers as if they had nothing to do with the weird constructions from the ZF axioms. Yet they take comfort in the *formal* equivalence of these constructions with the numbers that everyone uses. For this formal equivalence seems to show that one can go on using numbers, ordinary logic, and the prescribed operations on sets, without fear of contradiction. In effect, they treat set theory as a useful model, not as the context of their enterprise.

Thirdly, the actual role of set theory in mathematics has a number of curious aspects. It is something that everyone needs, because it provides a common vocabulary. So every mathematician has to learn the basic concepts and operations of set theory. But it's rather like learning a foreign language. One has to do it to function independently, but it's a tool and not an end. As a research endeavor, it's not very interesting, at least not to most mathematicians. So one learns the language and, in practice, applies set theory to mathematical domains or to subsets of mathematical domains, essentially as I have advocated and defended in this chapter. One continues, in general, to form valid mathematical concepts even if the *definitions* one offers are clothed in the unfortunate trappings of set theory.

Within their proper context, one can, as I have argued, justify the settheoretic operations that matter, and that the ZF axioms permit. On the other hand, notwithstanding the weird constructions of the set theorists, there is no actual need, for example, to combine existents of different kinds, as unions, into a single set.

In actual practice, once the real numbers are taken as a given, mathematicians observe hierarchy. And, to some extent, at

given, mathematicians observe hierarchy. And, to some extent, at least from a formal perspective, that observance is even reinforced by a set-theory that offers a different hierarchy of its own.

To a working mathematician, set theory offers the following:

- A concept that, in actual use, serves a very important function in mathematics, namely, the function of isolation that I detailed in the earlier part of this chapter,
- A uniform language and a short list of logical operations that, in an appropriate context, are valid and useful ways of expressing the logic in their mathematical analyses,
- A demarcation of permissible logical steps that avoids the contradictions that plagued the Cantor-inspired mathematics of the late 19^{th} and early 20^{th} centuries, ●
- A formal model that, however absurd, captures and reproduces the formal infrastructure of mathematics with a minimum number of formal axioms. That formal model, however inappropriately, is taken to provide a warrant for the formal soundness of their mathematical pursuits

the formal soundness of their mathematical pursuits.

In short, mathematicians take what they need, and ignore the rest.

General Conclusions

My goal in this chapter was to outline the proper role and scope of set theory in mathematics. I have discerned its essential function in isolating a portion of a mathematical domain, isolating a range of instances of a mathematical concept. In this, I have stressed the importance of hierarchy in mathematics. Concepts, pertaining to quantitative relationships in the world, come first. Sets presuppose mathematical abstraction; they presuppose a mathematical domain.

Sets provide differentiation within a domain, a domain consisting of systems of mathematical measurements or geometric objects of a particular type.

In discussing point set topology, I have provided an extended, nontrivial, example of the proper use of set theory to measure precision standards of convergence, an example that further indicates the ubiquity of geometry in mathematics, an example that also has great interest in its own right.

Finally, I have argued that the standard approach to set

theory starts in mid air, violates hierarchy, and has only a formal

connection to actual mathematics.

[1] Patrick Suppes, *Axiomatic Set Theory*, Dover Publications, Inc., New York, 1980, p 1-2

[2] Leo Corry, *Modern Algebra and the Rise of Mathematical Structures*, Birkhauser Verlag, Basel, 2004, Chapter 7, "Nicolas Bourbaki: Theory of Structures,"

p 289-338. Also, David Aubin, *Science in Context* 10, 2 (1997), "The Withering Immortality of Nicolas Bourbaki: A Cultural Connector at the Confluence of

Mathematics, Structuralism, and the Oulipo in France," pp 297-342

[3] Paul R. Halmos, *Naïve Set Theory*, Springer, New York, 1974, p 1

[4] J. Dieudonne, *Foundations of Modern Analysis*, Academic Press, New York, 1960, p 27-29

[5] David Hilbert, *Foundations of Geometry*, Open Court Classics, La Salle, Illinois, 1990, Chapter 1 "The Five Groups of Axioms," p 3–28

[6] Corry, p 323-334

[7] Mary Tiles, *The Philosophy of Set Theory*, Dover Publications, Inc. New York,1989, p 121-123. Also, Suppes. Also, Halmos

[8] Ayn Rand, *Introduction to Objectivist Epistemology Expanded Second Edition*, p 18 in the paperback edition

[9] Rene Descartes, *Des matiers de la Geometrie*, 1637, available in English translation as *The Geometry of Rene Descartes*, Dover Publications, Inc., 1954. Also, see Carl B. Boyer, *History of Analytic Geometry*, Dover Publications, Inc. Mineola, New York, 1956, Chapter V, p 74–102

[10] Rand, p 72-74, regarding borderline cases

[11] Rand, p 12

[12] Moritz Epple, Chapter 10 "The End of the Science of Quantity: Foundations of Analysis, 1860 – 1910," In *A History of Analysis*, edited by Hans Niels Jahnke (Rhode Island, American Mathematical Society, 2003 hardback) quoting Cantor on

p 312

[13] Pat Corvini, lecture entitled "Achilles, the Tortoise, and the Objectivity of Mathematics," summer, 2005, available on CD from the Ayn Rand Book Store

(www.aynrandbookstore.com), Corvini makes a related point in relation to the

meaning of mathematical infinity

[14] Boyer, *Analytic Geometry*, pp 74–102

[15] Euclid, *Elements*, edited with notes by Thomas L. Heath (New York: Dover Publications, 1956), Book 7

[16] Apollonius of Perga, *Treatise on Conic Sections*, Cambridge at the University Press, 1896

[17] Sir Thomas Heath, *Mathematics in Aristotle*, Thoemmes Press, Bristol, England, 1998, p 117

[18] Carl B. Boyer, *A History of Mathematics*, John Wiley & Sons, Inc., New York, 1991, p 31

[19] Boyer, *A History of Mathematics*, p 345 regarding Descartes rule of signs, p 411 regarding Newton's method

[20] Rand, p 18

[21] From a formal perspective, one can generate A union B from subsets of A × B, by taking the disjoint union A × b_0 and a_0 × B

[22] Harriman, "The Discovery of Universal Gravitation," in Chapter 4, "Newton's Integration"

[23] Herbert Goldstein,

Classical

Mechanics, Addison-Wesley Publishing

Company, Inc. Reading, Massachusetts, 1965, Chapter 1, p 1-29

[24] See Chapter 7

[25] See Suppes. Also, see Halmos

[26] Suppes, p 35, Halmos, p 24

[27] Suppes, p 86-89, Halmos, p 30-33

[28] On the power set axiom, see Suppes, p 46-48; Halmos, p 19-21. See, also, Penelope Maddy, *Realism in Mathematics* (Oxford, Clarendon Paperbacks, 1992), p 165 provides, as an example, "Two pairs of shoes are naturally viewed as a set of two

sets;" Also, regarding the scope of a set of subsets, p 102, "subcollections are

'combinatorially' determined, one for each possible way of selecting elements,

regardless of whether there is a specifiable rule for these selections. This is the

mathematical notion of a collection: a collection formed combinatorially, in a series

of stages that make up the iterative hierarchy."

[29] Jose Ferreiros, *Labyrinth of Thought*, Birkhauser, Basel,1999 provides an excellent history detailing how set theoretical concepts arose in mathematics

[30] Choquet, Gustave, *Topology*, Academic Press, New York, 1966, p 11

[31] Choquet, p 91

[32] Example from Choquet, p 91

[33] Choquet, p 92-93

[34] Choquet, p 97

[35] Halmos, p 2

[36] Halmos, p 4

[37] Ferreiros, *Labyrinth*

[38] Suppes, pp 1-2

[39] Epple, p 312 with reference to Cantor

[40] Epple, p 312

[41] Tiles, p 121

[42] Halmos, p 7, says, in reaction to Russell's paradox: "There is no universe."

[43] Epple, p 312

[44] Suppes, p 8-9

[45] Harry Binswanger, *Objectivist Forum*, Feb 1984 pp 13-14 for a related assessment of Russell's Paradox and Russell's theory of types. Binswanger writes:

"The actual solution to the paradoxes of self-reference lies in grasping that

statements are *cognitive* instruments. A statement, or proposition, is not just any seemingly grammatical collection of words; it is an integration of concepts used to

make a meaningful assertion … This means that no statement can refer *only* to itself."

[46] Epple, p 311

[47] Suppes, p 5

[48] Halmos, p 4-6

[49] Suppes, p 11

[50] Tiles, p 121

[51] Tiles, p 122

[52] Tiles, p 121

[53] Tiles, p 124

[54] Halmos, p 44

[55] Maddy, p 84, identifies two standard sequences of sets that have been identified with the natural numbers, namely the von Neumann ordinals and

Zermelo's ordinals. On this basis, p. 85, she writes, "If one of these particular choices is the correct one, that is, if one sequence of sets really is the numbers, then there ought to be arguments that tell us which sequence that is. … But there are no such

arguments. Therefore, numbers are not sets." Later, on the same page, "Therefore,

numbers are not objects at all."

[56] But see previous footnote (Maddy)

[57] Tiles, p 123, states the power set axiom

[58] Maddy, p 117-125, discusses related issues such as the axiom of choice, Cohen's finding on the continuum hypothesis, and the search for additional axioms

to complete ZF axioms

[59] Epple, p 321

[60] Tiles, p 95. Also, see David Hilbert in Paul Benacerraf and Hilary Putnam, *Philosophy of Mathematics Selected Readings* (New Jersey, Prentice Hall, 1964

hardback), "On the Infinite," (1926), p 141, "No one shall drive us out of the paradise which Cantor has created for us."

Chapter 7 Vector Spaces: A Study in Mathematical Abstraction

Introduction

The nineteenth century was the most eventful century in the entire history of mathematics. In the conventional view, the nineteenth century movement toward greater mathematical rigor and abstraction simultaneously discarded the geometric foundations of mathematics and even its mission, across millennia, as the science of quantity.

The nineteenth century changed the practice of mathematics and it also changed the way that people *think* about mathematics. Yet, contrary to the modern, almost universal, consensus, the movement to greater mathematical abstraction and rigor that began in the late nineteenth century did not change, and has not changed, the *fundamental nature* of mathematics. It remains, even as it is practiced, in Ayn Rand's characterization, "the science of measurement."[1]

I say this, despite the formal abandonment of any relationship of mathematics to the world and despite the rise and later evolution of set theory, culminating with the ZermeloFraenkel axioms of set theory in the twentieth century, discussed in Chapter 6.

But to show what I mean by my claim will require more than discussions of elementary mathematics. For I maintain that the many abstract mathematical sub-specialties, in, for example, geometry, differential equations, and abstract algebra, that flourished in the twentieth century, remain, in their method and in their achievements, though not in their presentations, vigorous sub-specialties of the science of measurement.

A full vindication of this viewpoint would require a broad survey of the conceptual foundations and methods of the major branches of $20^{th}/21^{rst}$ mathematics. I can only *begin* that task within the scope of this book.

As a first example, I explained, in Chapter 6, how the fundamental definition of a topological space serves to address a very general problem of measurement,

namely the need to measure processes of successive approximation. I will devote this chapter and the next to showing how the measurement perspectives I have presented in this book illuminate and provide a similar conceptual foundation for two other important domains in modern mathematics. My treatment of these subjects will be elementary, but the subjects are both more advanced and more abstract than those treated in the first five chapters of this book. Both subjects, vector spaces and abstract groups, have fundamental importance in both their scientific and mathematical applications.

This chapter explains the most fundamental concepts of *vector spaces*. Vector spaces have a central, a fundamental, and a pervasive importance in mathematics. The theory of vector spaces is part of abstract algebra. And yet, in its mathematical and scientific applications, the analysis of vector spaces retains a form that is reasonably concrete.

Historically, the theory of vector spaces is a bridge from the prior, more concrete, perspectives in mathematics to the more abstract modern perspectives. In this book, the theory of *vector spaces*, also known as *linear algebra*, will serve as a similar bridge in understanding the universal role of measurement in mathematics. For this reason, vector spaces provide a useful case study in mathematical abstraction.

The modern concept of *vector space*, and of the *linear transformations* that relate them, is a culmination of a development across millennia. The implicit concept of a vector is at least as old as Archimedes.[2] Linear algebra, the study of simultaneous linear equations is even older: The Babylonians were solving simultaneous linear equations in two unknowns 4,000 years ago.[3]

In its earliest physical form, a vector is a magnitude associated with a particular direction. A displacement is a vector; so is a force.

But the need for solving simultaneous equations is far broader than the need for the physical concept of a vector. Simultaneous equations arise as a problem in indirect measurement. They arise whenever two or more relationships are known about an unknown quantity or a set of quantities.

If I say, for example, that John is three years older than Mary, that the sum of their ages is 17, and I then ask for their respective ages, I am posing a problem that involves simultaneous equations. Whenever Euclid intersects two lines, he is solving, geometrically, two simultaneous equations.

solving, geometrically, two simultaneous equations.

The connection between these perspectives, the connection of the algebra to Euclid, is supplied by analytic geometry. One expresses the equations for a pair of lines as, for example:

$x + y = 17$ $x – y = 3$

One solves these equations to find their intersection, concluding that $x = 10$ and $y = 7$.
The modern theory of linear algebra integrates the physical theory of vectors with the older theory of simultaneous linear equations. Like analytic geometry, it brings analytic methods to bear on geometric problems and provides a geometric perspective on problems that have no direct relationship to space and time.
The value of and need for vector spaces and linear algebra should not be controversial. As a fairly early example of the modern abstract approach, it is also ideal for my purposes, because of the way that the concepts of linear algebra further exemplify the relationship between quantity and the measurements by which quantities are quantified.

Vector Spaces

In its most primitive form, vectors date back to Archimedes. The vectors of Archimedes are forces, forces acting in but two directions: up and down. In one of his analyses, the forces are weights acting on lever arms. In another, Archimedes starts with the understanding that objects float when their gravitational force is balanced by the opposing force of water pressure. Archimedes' analyses are analyses of forces in balance.

Archimedes discovered his famous law of levers by analyzing centers of gravity. He starts from a simple empirical observation: Two equal weights standing at equal distances from a fulcrum will balance. Moreover, the force exerted on the fulcrum is equal to the combined force of the two weights and is, therefore opposed by an equal force exerted by the fulcrum. This is illustrated in the following diagram:[4]

Figure 1

From this observation, Archimedes either concludes or, perhaps, also observes that two equal weights located elsewhere on a balance beam will, likewise balance an opposing force at *their* midpoint, no matter where the fulcrum of the entire beam might be located. That is, the equilibrium of the beam will not be changed if the two equal weights are each moved the same distance in opposiste directions. It is as though a second scale were located at their midpoint. This situation is illustrated in Figure 2 as it relates to Figure1:

Figure 2

In this diagram the two equal weights on the right have been moved equal distances in opposite directions from their original positions in Figure 1. Regardless of how far each weight is moved, as long as these distances are equal, the scale will always remain in balance.[5] In the form needed by Archimedes, if two equal weights are each moved to their midpoint on a balance beam, this movement will not affect the balance.

However, although this principle is assumed in Archimedes' argument, he neither states that principle as a premise (or observation) nor provides a demonstration. But what he offers, with full rigor thereafter from this base, is a beautiful demonstration of his famous law of levers:

"Two magnitudes, whether commensurable or incommensurable, balance at distances reciprocally proportional to the magnitudes."[6]

Archimedes' argument is stated, and proven, in a very general way, but his approach to demonstrating his law amounts to the following:

Suppose one wants to balance three equal weights against five. Start with the

following scale, in which eight weights (eight being the sum of three and five) are equally spaced on a balance beam. The beam will balance when the fulcrum is placed in the middle:

Figure 3

As a first step toward balancing 3 of these weights against 5, Archimedes, in effect, consolidates the three weights on the right, without disturbing the balance of the beam, as follows:

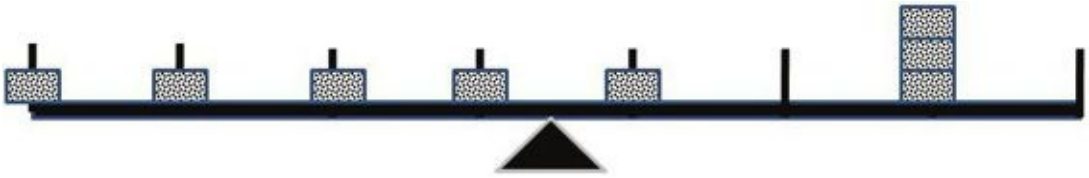


Figure 4

The balance is maintained because the two weights were both moved the same distance in opposite directions.
Next, in the same fashion, he consolidates the five remaining to their central point, again moving weights, in pairs, for equal distances in opposite directions:

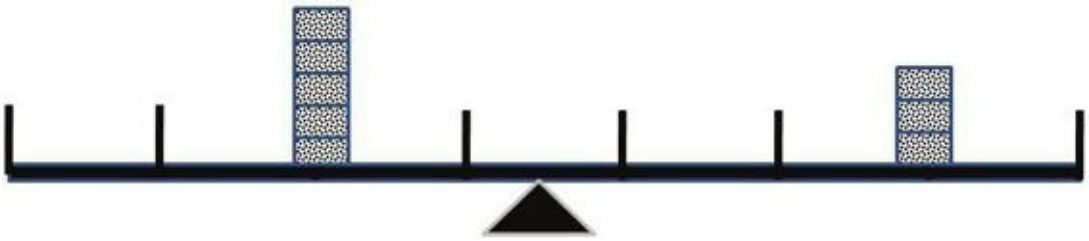


Figure 5

Once again, for the same reason, the scale remains in balance.
It remains to relate the respective distances from the fulcurm of the two piles of weights. These respective distances from the fulcrum are 1.5 units and 2.5 units. One can verify, by multiplying, that the product of the weight on the left (5) times the distance from the fulcrum (1.5) is equal to the weight on the right (3) times its distance from the fulcrum (2.5). Numerically, both products are 7.5. This relationship will always hold: the scale will balance if and only if the product of weight and distance on the left is equal to the corresponding product on the right. In the modern formulation, these two products are called *moments*, and the Law of Archimedes states that the left moment is equal to the right moment.

Archimedes, however, states his law a little differently, as an equality of two ratios. The corresponding weights are 5 and 3. In terms of these numbers, Archimedes law of levers states, and one verifies in this case, that 2.5/1.5 = 5/3. When the beam is in balance, the distances of the weights from the fulcrum are in inverse proportion to the weights. From a modern perspective, the modern formulation is mathematically equivalent to that of Archimedes. But, as noted in Chapter 2, the Greek mathematicians did not consider products of magnitudes.
One can apply this approach to any number of weights. For example, to balance 5 equal weights against 7 equal weights, start with 12 (= 5 + 7) equally spaced weights and consolidate 5 on the right and 7 on the left. One finds, again, the inverse proportion of Archimedes and the equality of left and right moments of the modern formulation.
One could translate this process into a general algebraic argument to establish the general case. However, for technical reasons, a slight variation of the argument is somewhat easier to follow.
So suppose that weights of m × W and n × W are spread out side by side along the balance beam. Here, m and n are integers and W is taken as the amount of the total weight spanning any two successive vertical distance markers. So the weight m × W spans m distance units and n × W spans n distance units.The situation is illustrated in Figure 6:

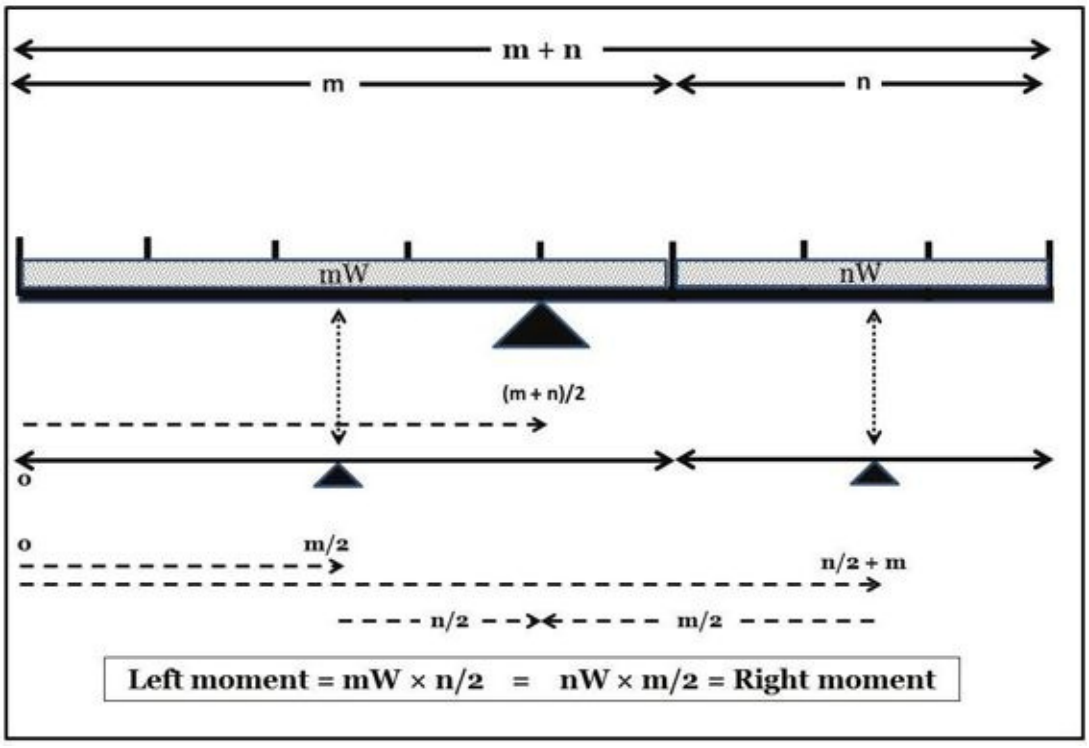


Figure 6

Figure 6

The two green bands represent the two weights mW and nW. (I now omit the times sign, × between the integer and the weight.). Since the entire length of the balance beam spans m + n units of length, the midpoint at which the beam balances is at the point (m + n)/2 units from the leftmost edge of the beam. Similarly, the midpoint of the portion of the beam occupied by mW, on the left, is at m/2 units and the midpoint of the portion occupied by nW, on the right, is at n/2 units past the end of mW, and, therefore, at the point n/2 + m length units from the leftmost edge of the beam.

One checks that the midpoint of the left mW weight, in relation to the center of the beam, is a distance of n/2 (= (m + n)/2 – m/2) units from the center of the beam and the right midpoint of the nW weight is a distance of (n/2 + m) – (m + n)/2 = m/2 units from the center of the beam. Then as stated in Figure 6, one notices that mW × n/2 (left moment) = nW × m/2 (right moment) = nmW/2. This is the Law of Ar*chimedes stated as an equality of moments.*

The consolidation of each weight at its respective center is illustrated in Figure 7:

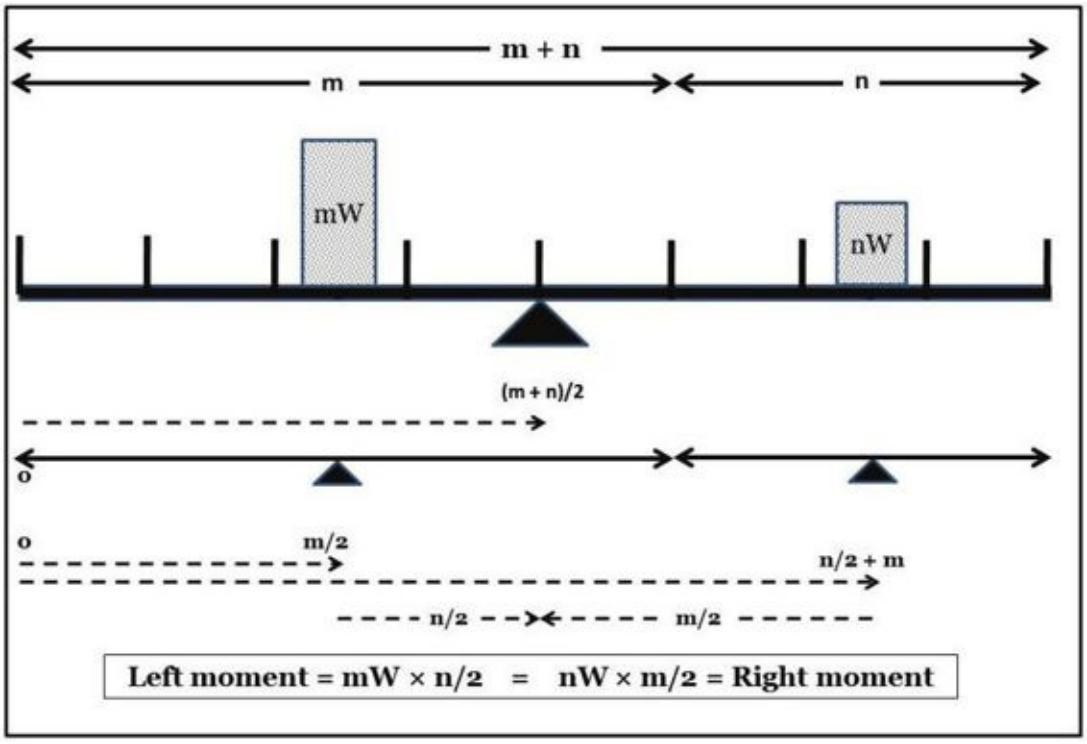


Figure 7

Archimedes' proof is more elegant, though less intuitive, than this rendition, but this is the central idea behind his method. His argument proves his law in

general, for any pair of commensurate weights. He then applies the Eudoxus conception of ratio to extend his conclusion to incommensurable magnitudes and, thereby, to complete his argument.[7]

The study of forces in equilibrium and the first seeds of the mathematical concept of a vector begin with this proposition.
I said in Chapter 2 that Archimedes used line segments to represent force. And so he does. But his forces are not just magnitudes; they act in a certain direction, even if that direction is either up or down.
This directionality of force is especially evident in Archimedes's analysis of floating bodies. To state his key propositions:

"Any solid lighter than a fluid will, if placed on the fluid, be so far immersed that the weight of the solid will be equal to the weight of the fluid displaced."[8]

and
"If a solid lighter than a fluid be forcibly immersed in it, the solid will be driven upwards by a force equal to the difference between its weight and the weight of the fluid displaced."[9]

Archimedes also states a proportion:

"If a solid lighter than a fluid be at rest in it, the weight of the solid will be to that of the same volume of the fluid as the immersed portion of the solid is to the whole."[10]

Archimedes's understanding of buoyancy reflects a focus on one key question: At what point does the upward force exerted by water pressure balance the weight of the object?

Archimedes's analysis of forces is restricted to one dimension. But his treatments of levers and buoyancy are the first analyses in history of the equilibrium of forces, of the combined effect of various forces acting on a body.

In the modern era, one of Newton's great innovations was to treat velocity and acceleration as vectors, as magnitudes associated with a particular direction. And, further, to relate both velocity and acceleration to forces, magnitudes acting in a particular direction. Newton's innovation was essential to the discovery of his laws of motion.[11]

Vectors are first conceived geometrically, but they can be meaningfully related arithmetically in a way that has both geometric and physical significance. Most importantly, two vectors of the same type can be added to yield a third vector of that same type.

The easiest way to understand how vectors add is to start with vectors of displacement. Suppose one goes one mile east and two miles north. In coordinates, one goes from the point (0, 0), one's starting place, to the point (1, 2). Next, one goes 3 miles east and 1 miles north. In all, one has gone 4 miles east (1 + 3) and 3 miles north (2 + 1). The second displacement, had it acted from the origin, would be written as (3, 1). And, as a *displacement*, the relevant characteristics are, in fact these very coordinates. One doesn't measure a *displacement* by where the displaced object ends, but by the *change* in its coordinates during the displacement.

The total displacement in each direction is the sum of the displacements, of the net changes in each of the respective directions. In other words, the total displacement is given by (1, 2) + (3, 1) = (4, 3): The combined effect of two displacements is found by adding their coordinates. Figure 8 depicts these relationships:

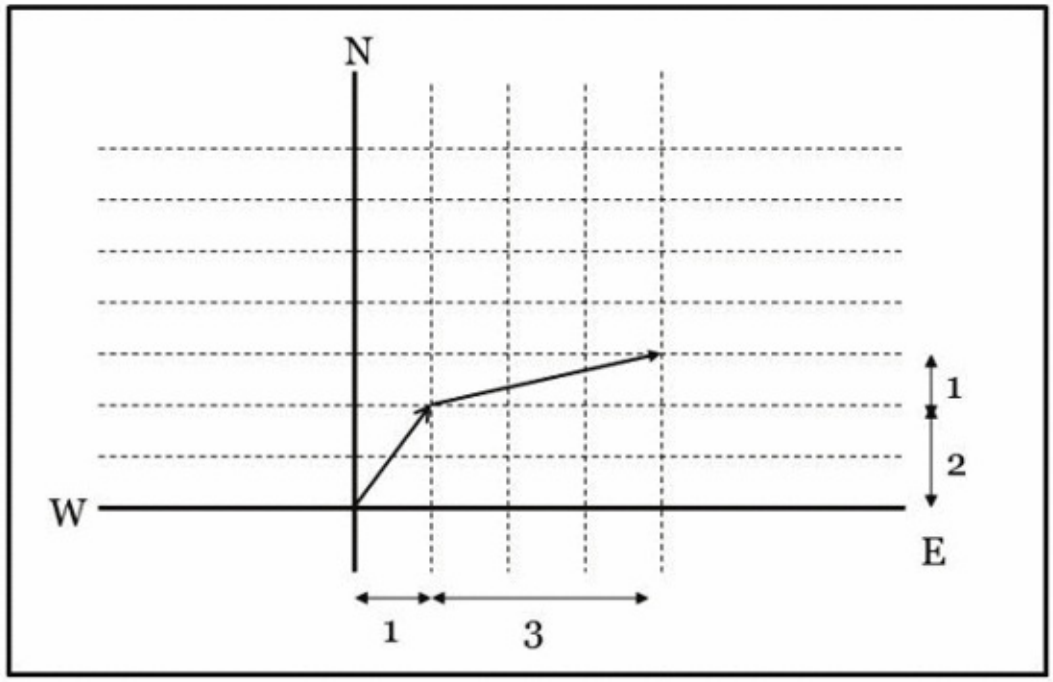


Figure 8

The reverse displacement has the same magnitudes, but in the opposite direction and their combined effect is the zero displacement (0, 0). So, for example, (1, 2)

and their combined effect is the zero displacement (0, 0). So, for example, (1, 2) + (-1, -2) = (0, 0). It's also clear that the order in which the two displacements are carried out does not affect the outcome. In short, adding works for displacements, the same way it works for numbers. One would not call it addition, even in this generalized sense, were that not the case.

But the relationship, between addition of displacements and ordinary addition, is even stronger than that. Each direction, north and east, represents a distinct axis. Progress along either of these axes is independent from progress along the other axis. What I'm calling the sum of two displacements can, with perhaps even greater justice, be characterized as simply carrying out two additions, additions of simple magnitudes, pertaining to two separate measurements of the moving object. To bring them together in one expression, to take them together as a displacement, is to make an intellectual integration. But it is the integration of two independent measurements, of two independent axes of measurement, into one more complicated measurement.

What I have discussed for displacements, applies equally well to velocities, accelerations, and force. A displacement is a magnitude, a distance moved, in a certain direction, the direction of the displacement. A velocity is a speed in a certain direction, i.e., a displacement per unit time in a certain direction. Acceleration is the rate of change in a velocity; that rate of change resolves itself as a magnitude in a certain direction. Indeed, the coordinates that measure the acceleration of a moving object are calculated as the time derivative of the respective coordinates of its velocity.

Finally, for forces, one finds that the vector sum of two forces has the same physical effect as the two forces acting independently. This is usually referred to as the parallelogram law of forces and the equivalence of that law to the coordinate conception is illustrated in the following elaboration of Figure 8:

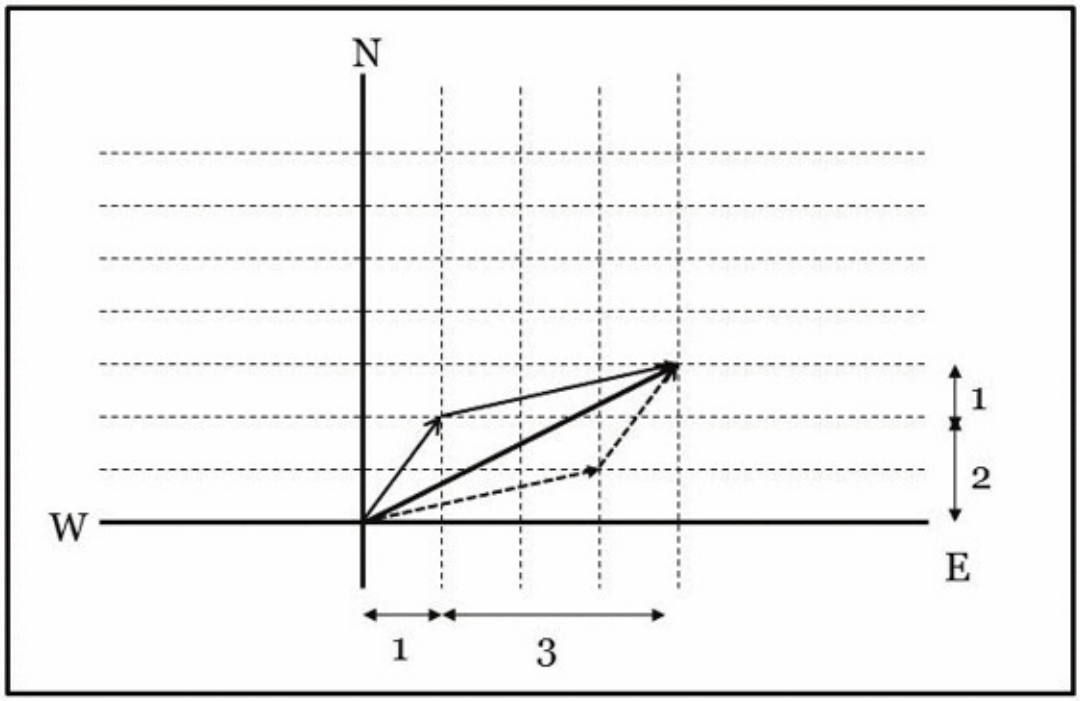


Figure 9

In Figure 9 one sees that, if the vector operations are performed in reverse order, the resulting picture, illustrating both sequences of displacements, is a parallelogram and the vector sum, of either sequence, is the diagonal of the parallelogram.

From a measurement perspective, one uses Cartesian coordinates. But this is a *choice* one makes in how one measures. From a physical perspective, a displacement or a velocity is a unitary fact, albeit a fact with distinguishable aspects. To even look at a velocity as a speed in a certain direction involves a choice, probably the simplest, among reasonable perspectives. Yet, the complexity involved in this last choice resurfaces when one measures direction.

So far, a vector is a magnitude that is also directional. Expressed in coordinates, one adds vectors by adding the respective coordinates.

Magnitudes can be multiplied by numbers. Three miles per hour in a south-eastern direction multiplied by ten is 30 miles per hour in a south-eastern direction. If expressed in coordinates, to multiply the vector by 7 is to multiply each coordinate by 7. And this, just because each coordinate represents a separate, independent measurement. Thus, $7 \times (1, 2) = (7, 14)$.

Not to belabor the point, but multiplication interacts with addition in the usual way. Specifically, the so-called distributive law holds:

way. Specifically, the so-called distributive law holds:

$5 \times ((1, 2) + (3, 5)) = 5 \times (1, 2) + 5 \times (3, 5)$
This can be seen by calculating each side of the equality, comparing the results, and realizing why the results are equal:
$5 \times (1, 2) + 5\times (3, 5) = (5, 10) + (15, 25) = (20, 35)$ $5 \times ((1, 2) + (3, 5)) = 5 \times (4, 7) = (20, 35)$

As the calculation illustrates, the distributive law holds for vectors because it holds, separately, for each coordinate: for each independent measurement that one treats as a separate coordinate.

Vectors can be drawn, as arrows, from the origin in the Cartesian plane. For every point in the Cartesian plane there is a vector ending at that point. Equipped with this addition of vectors and with multiplication by numbers, the Cartesian plane can be regarded as a vector space. And there is nothing special about the number of coordinates. Exactly the same approach to addition and to multiplication by numbers works for 3 coordinates, for 4 coordinates, and, in general, for n coordinates where n is any integer greater than 0. With these definitions of addition and of multiplication by numbers, R^n can be regarded as a vector space.

But addition of vectors and multiplication by numbers is not restricted to coordinate pairs or even to n-tuples of coordinates representing points or vectors in R^n.

For example, as I already illustrated in Chapter 6, one can add polynomials and multiply them by numbers, as well. For example,

$(x^2 + x – 5) + (2x^2 – 2x + 7) = 3x^2 – x + 2$ and $3 \times (x^2 + x – 5) = 3x^2 +3 x – 15$

Addition of polynomials and multiplication by numbers for polynomials obeys the same distributive law that they do for numbers, just as they do for coordinate pairs.

And, in general, one can say the same for functions, as long as these functions take values that can be added and can be multiplied by numbers. To repeat the formulas from Chapter 6, one adds functions f and g, by adding their values at every point:

$(f + g)(x) = f(x) + g(x)$

$(f + g)(x) = f(x) + g(x)$
One multiplies a function by a number A by multiplying the value of the function at every point:
$(Af)(x) = A \times f(x)$

Read this as: The value of the function *Af*, at a point x, is equal to A times the value of the function f at the point x.
Addition of functions works the way it does for numbers. Multiplication by numbers and addition of functions working together, obey the distributive law, simply because they obey it at every point. Adding commutes with multiplication. One can, without difference in effect, carry out the multiplications first and then add the results, or one can add the functions together first and then apply the multiplier.
One can even look at coordinate vectors and functions in a similar way. From a certain perspective, one can view an n-tuple of coordinates as a function on the set of integers from 1 to n. Namely, if the coordinates are $(y_1, y_2, \dots, y_n)$[12] then the related function maps the number i to the coordinate y_i. To make this more concrete, one looks at the 3-tuple (2, 4, 5) as a function *f* acting on the domain {1 , 2, 3} by the rule f(1) = 2, f(2) = 4, and f(3) = 5.
And one can turn this around. Speaking somewhat loosely, think of a function f(x) as having an infinite number of coordinates, all them particular values of x contained in the domain of the function. From this perspective, the value of the function at x, namely f(x) is thought of as the xth coordinate. In adding two functions, one performs an independent addition for each value, x, of the independent variable. In multiplying a function by a number, one performs, in effect, a separate multiplication for each value of x. The laws for addition of functions and for multiplying them be numbers follow from the laws of arithmetic.
But this point is somewhat of a diversion. In regards to their abstract status as vectors, the important point is this: From a formal perspective, n-tuples of numbers, functions, and polynomials (a special class of functions) are all subject to the same kinds of mathematical operations. They all have an addition operation that functions the way addition is supposed to function. They can all be multiplied meaningfully by numbers and the multiplication obeys the distributive law with respect to the addition operation. And as mathematical domains, they all have a certain closure property. If one adds two of anything in the domain, one gets something else in the domain. If one multiplies something in the domain by a number the result is something in the domain. There is a zero value in the domain (the result of multiplying anything in that domain by zero).

And the negative of anything in the domain (the result of multiplying by -1) is also in the domain.

Any mathematical domain satisfying these characteristics is called a vector space.

Indeed, we studied such a domain in Chapter 2. When I discussed the pre-arithmetic of magnitudes, I was talking about one-dimensional vector spaces, except that I usually ignored negative magnitudes.

In a certain sense, mathematicians maintain the distinctions among these different kinds of vector spaces. Vector spaces, as mathematical domains, are regarded abstractly, but even on an abstract level, one regards them as distinguishable, as possibly referring, in *application*, to different concretes. These distinctions become particularly relevant when one introduces other measurements into the mix: such measurements, for example, as inner products and topological structures. I will discuss inner products later in this chapter; I discussed various topologies involving function spaces in Chapter 6.

But there is a great body of knowledge about vector spaces that doesn't depend on these other facets. And a systematic treatment of this knowledge, though always illuminated by particular examples, treats these other facets as omitted measurements.

The abstract theory of vector spaces is a theory that embraces the things that all vector spaces have in common. Mathematicians bought into this abstract approach for a very good reason: because otherwise they would find themselves proving essentially the same theorems, by essentially identical arguments, in special case after special case. Avoiding such redundant efforts, achieving what Ayn Rand calls unit economy,[13] and integrating one's knowledge is what abstractions are for.

One can be led, or misled, to believe that to think about something abstractly, especially in mathematics, is to think of it as being nothing in particular. Perhaps, for example, one might view abstract thinking in mathematics as a formal game, as following a prescription for manipulating meaningless symbols. But this is simply a perversion, a misunderstanding, of what abstractions are. Abstractions, as mathematical history and practice illustrates, are a means of conceptual integration. When one thinks about vector spaces, one is thinking about R^n *and* polynomials *and* functions, thinking about them from a perspective that applies to all of them without regard for their irrelevant differences. It is a perspective that focuses on the *structure* of a complex of relationships and omits from view the distinctions among the embodiments of that structure.

The term "abstract vector space", when it is used, is really a redundancy. An

abstract vector space is really just a vector space, not a separate concept or floating abstraction. When one proves theorems about vectors spaces, as such, one proves theorems that do not depend on the differences between polynomials and vectors in R^n; these theorems apply to both equally and in the same respects. For example, the configuration space of potential velocities and accelerations of a *particular* object in the universe is a vector space. A particular velocity of a particular object is a vector. We saw, in Chapter 2, that the concept of *magnitude* applies universally, to attributes such as mass and volume pertaining to any existing object, any future object, or any potential object in the universe. Similarly, every magnitude with a directional sense, attributable to any object in the universe, is a *vector*.

In dealing with vector spaces abstractly, one does not, thereby, wipe out their referential character. Rather, one treats the particular quantities to which they apply as omitted measurements. And, in so doing, in comparing different vector spaces, one focuses on the *structural* similarities and differences between the vector spaces.[14]

In this regard, compare vectors in R^{33} tuples of numbers (a, b, c), with quadratic polynomials $ax^2 + bx + c$. A 3-tuple is determined by three *coordinate numbers*, a, b, and c. A quadratic polynomial is determined by its *coefficients* a, b, and c. What happens if I add two 3-tuples? As a representative example, for 3-tuples,

$(1, -2, 3) + (3, 2, -1) = (4, 0, 2)$

For polynomials,

$(x^2 – 2x + 3) + (3x^2 + 2x -1) = 4x^2 + 2 = 4x^2 + 0x + 2$

The behavior of the coordinates, in the first case, exactly matches the behavior of the coefficients in the second. What about multiplication? For 3-tuples,

$4 \times (1, -2, 3) = (4, -8, 12)$

and, for polynomials,

$4 \times (x^2 – 2x + 3) = 4x^2 – 8x + 12$

Again, the behavior matches exactly.

One can map pairs of 3-tuples to polynomials, add them together or multiply them by numbers, and then reverse-map the resulting polynomials back to 3-tuples. The resulting calculations are unaffected by the detour into the polynomial domain. In the above example, one maps (1, -2, 3) and (3, 2, -1) to $(x^2 – 2x + 3)$ and $(3x^2 + 2x -1)$. Adding these polynomials, one then maps the

result, $4x^2 + 0x + 2$, back to (4, 0, 2). Which we've already seen to be the sum of (1, -2, 3) and (3, 2, -1).

As far as the kinds of calculations that apply to all vector spaces, addition of vectors and multiplication of vectors by numbers, are concerned, there is no difference in the arithmetic of 3-tuples and the arithmetic of quadratic polynomials. There is an exact correspondence. As vector spaces, one says that they are *isomorphic* and the map, or correspondence, that I exhibited between them is, regarded as a map, an *isomorphism*. As far as their *vector-space structure* is concerned, they are identical.

As systems of measurements, n-tuples and polynomials of degree (n – 1) correspond to different things in the world, just as magnitudes of length versus magnitudes of weight correspond to different things in the world. And when it matters, for whatever reason, one focuses on these differences. But when one studies vectors abstractly, one omits these differences, because the conclusions and understandings one is reaching don't depend on these differences.

Structurally, as vector spaces, one regards n-tuples and polynomials of degree (n – 1) as identical, on the principle that each vector space must measure *some* relevant aspect of the world, but may measure *any*. In any concrete case, specific vectors must relate to *some* appropriate measurable concrete, but may relate to *any*. This relationship to the world might be more direct in the case, say, of 3-tuples than in the case of quadratic polynomials. But, from the standpoint of their vector space structure, this doesn't matter. All of these considerations, from an abstract perspective on vector spaces, are omitted measurements.

One treats isomorphic vector spaces as identical structurally. But one also distinguishes them when necessary or appropriate. In particular, when one maps vectors from one vector space to another, one treats the two vector spaces, thus related, as *separate and distinguishable*, just as they would actually be in application to concretes. But one does this, while, at the same time, regarding any *particular* application to concretes as omitted measurements.

For example, Newton's famous formula F = ma relates two different vectors, one representing the force on an object and the other representing the acceleration that the force causes. Both the force and the acceleration are thought of, mathematically, as vectors in R^3. But the vector space of forces is a different vector space than the vector space of accelerations. The two instances of R^3 are treated, properly, as two separate, though structurally identical, mathematical domains.

As a concrete quasi-numerical example, consider the map A from R^2 to R^2 defined as

A(a, b) = (2a, 3b)
Applying the formula to the coordinate pair (10, 5):
A(10, 5) = (20, 15)

In general, the vectors in the *domain* of the mapping A will represent a different kind of quantity than the vectors in the *range*, i.e., in the second vector space, just as in the example from Newton's physics. So, even though one represents both vectors as coordinate pairs, one, nonetheless, treats them, with or without comment, as *distinct*. And this is certainly the case with the isomorphism I exhibited between 3-tuples and quadratic polynomials.

In this, mathematicians are acting on the right principle. And, in regards to abstraction in general, when mathematicians give examples of mathematical concepts, they are implicitly acknowledging that these examples are what the concepts are about, are referents of the concepts. Finally, the vector space concept, in *application*, is open-ended or as I put it in Chapter 6, wide-open ended. Considered structurally, one considers isomorphic vector spaces as identical, but considered in application, there are no known limits to the applicability of a vector space of a given structure. A vector space, considered abstractly, without regard to a specific application to external referents, considered as a specific range of mathematical possibilities, as a mathematical domain, is a set. But, insofar as one distinguishes, say, force vectors from acceleration vectors, there is no such thing as the set of all vector spaces.

In sum, mathematicians view two instances of R^3 as the same, but also as different, though not in the same respect. They are the same structurally but are distinct instances, distinct mathematical domains that have the same vector-space structure.

When one looks at vector spaces in this way, one treats them, not as systems of *measurements*, but as systems of *quantities*, as an abstract perspective on something in the universe: potential velocities or accelerations of an object or possible forces acting on an object. One's perspective is *geometric*. Vectors, as such, are not dimensionless ratios the way that numbers are. Vectors are not a *measurement* of an object or attribute; they are the object or attribute that is being measured. They are measurable aspects of an object considered as something external that is measured and considered solely in relation to a set of measurements. One looks at vectors the way that I looked at magnitudes in my treatment of the pre-arithmetic of magnitudes in Chapter 2.

And this is true, at one remove, even when the vectors are polynomials. It is true because the essence of the geometric perspective is its *focus on an object*, any object, even if that object itself is a mathematical abstraction. The geometric perspective is a focus on something that is taken to have an independent existence, as being an object of awareness. It is true that polynomials are the product of an earlier action of consciousness. But, as an object of further study, they exist prior to that subsequent study.[15]

The measurement side arises when one establishes coordinates, chooses coordinates having an unspecified, but, in any concrete instance, specifiable relationship to concretes. N-tuples, considered as ordered sets of numbers applying to some corresponding, unspecified set of measurements, are a system of measurements. A vector space considered geometrically, by contrast, is a system of quantities, a system that contemplates actual or potential existents, bearing determinant relationships to each other, relationships that *are independent* of any particular system of measurements, relationships that *can be studied independently* of any particular system of measurements.

Polynomials are an interesting case because, as such, qua polynomials, they are a system of measurements. Nonetheless, when they are considered as elements of a vector space, these elements are quantities having measurable relationships to each other. And, of course, as in the case of numbers, these relationships derive from the relationships between the things that polynomials measure.

A coordinate space, R^n, regarded as a vector space is a similar case in point. A coordinate space, as such, constitutes a system of measurements, measurements, indeed, relating to a vector space. But the particular constellation of magnitudes, and the units by which they are to be measured, are, once again, omitted measurements. If one focuses on the arithmetic of a coordinate space then, from that perspective the coordinate space is a system of measurements. But if one focuses on these n-tuples as referring to a system of unspecified quantities, as one necessarily does when considering the effect of coordinate-system changes, then, from that geometric perspective, they are a system of quantities.

Indeed, the relationship of the two perspectives, here, is similar to the relationship between the real numbers, considered as a *system of measurements* and the real number *line*, considered as something *measurable*. To pursue the analogy a little further, in the first instance, numbers are looked at measurements of something external whereas, in the second instance, they are looked at as a

range of external possibilities, considered in measurable relationships to each other.

Bases and Dimension

A vector in R^3 is uniquely specified as a 3-tuple (a, b, c). One thinks of the first coordinate as a value in one direction, say the x direction and the second as the value of the vector in a second direction, namely the y direction. One way to make this relationship more explicit is to write:

(a, b, c) = a × (1, 0, 0) + b × (0, 1, 0) + c × (0, 0, 1)
(Here, '×' refers to multiplication.)

Clearly, every vector in R^3 can be uniquely represented in this form. In this context, one calls the three vectors, (1, 0, 0), (0, 1, 0), and (0, 0, 1), *basis* vectors and considers the three vectors, taken together, as constituting a *basis* for the vector space.[16] One often introduces special symbols for these vectors. For example, one may write:

$(1, 0, 0) = \hat{e}_1$
$(0, 1, 0) = \hat{e}_2$
$(0, 0, 1) = \hat{e}_3$

In this notation,
$(a, b, c) = a\hat{e}_1 + b\hat{e}_2 + c\hat{e}_3$

This notation indicates the same relationship of a vector to its basis, but with less clutter, using fewer symbols. Moreover, this means of expression is equally applicable to other vector spaces. For example, a natural basis for quadratic polynomials is the following:

$x^2 = \hat{e}_1$ $x = \hat{e}_2$ $1 = \hat{e}_1$

In this table, all three expressions, x^2, x, and 1 should be regarded as quadratic polynomials. For example, one should regard 1 as the quadratic polynomial $0x^2 + 0x + 1$. One has, in this notation,

$ax^2 + bx + c = a\hat{e}_1 + b\hat{e}_2 + c\hat{e}_3$

There is, however, nothing unique about these particular bases. Just as one can change coordinates in the Cartesian plane, one can change bases in a vector space. The reason for changing bases is less obvious for polynomials, but, even in that case, there are situations in which a different basis, a basis consisting of polynomials that satisfy some additional criterion, is more appropriate.

As an example of a change in basis, consider the three vectors:

$\mathbf{v}_1 = (1, -1, 0)$ $\mathbf{v}_2 = (1, 1, 1)$ $\mathbf{v}_3 = (0, 0, 1)$

I say that any vector (a, b, c) can be written as the following linear combination of $\mathbf{v}_1$, $\mathbf{v}_2$, and $\mathbf{v}_3$: Namely,
(a, b, c)
$= (a/2 – b/2)\mathbf{v}_1 + (a/2 + b/2)\mathbf{v}_2 + (c – a/2 – b/2)\mathbf{v}_3$
For example, $(2, 4, 5) = \mathbf{v}_1 + 3\mathbf{v}_2 + 2\mathbf{v}_3$.
One checks the general formula by the following computation:

$(a/2 – b/2)\ \mathbf{v}_1 + (a/2 + b/2)\mathbf{v}_2 + (c – a/2 – b/2)\mathbf{v}_3 = (a/2 – b/2)(1, -1, 0) + (a/2 + b/2)(1, 1, 1) + (c – a/2 – b/2)\ (0, 0, 1)$
$= (a/2 – b/2, -a/2 + b/2, 0)$
$+ (a/2 + b/2, a/2 + b/2, a/2 + b/2)$
$+ (0, 0, c – a/2 – b/2) = (a, b, c)$

But how did I discover this particular relationship to begin with? I set up and solved a system of simultaneous equations, thus:
First write down the problem one needs to solve. One is looking for coefficients A, B, and C such that

$A\mathbf{v}_1 + B\mathbf{v}_2 + C\mathbf{v}_3 = (a, b, c)$
Expanding this relationship,
$(A, -A, 0) + (B, B, B) + (0, 0, C) = (A + B, -A + B, B + C) = (a, b, c)$ Equating corresponding coordinates leads to the system of simultaneous linear equations:

$A + B = a$
$-A + B = b$ $B + C = c$

One solves this system of equations by standard techniques to derive the

expression of A, B, and C in terms of a, b, and c.
In general, to change bases in a vector space, to translate a vector expressed in terms of one basis to an expression in terms of a different basis, requires solving such a system of simultaneous equations.
The equations, in this case, had a unique solution. They can be solved for any set of constants a, b, and c and there is only one set of values, A, B, and C that simultaneously satisfies all three equations.
But such is not always the case. For example, suppose I had, instead, chosen the vectors:

$\mathbf{v}_1 = (1, -1, 0)$ $\mathbf{v}_2 = (1, 2, 1)$ $\mathbf{v}_3 = (3, 0, 1)$

Now I need to solve the equations:

$A + B + 3C = a$
$-A + 2B = b$ $B + C = c$

Adding the first equation to the second yields:
$3B + 3C = a + b$
Next, subtracting 3 times the last equation from this one yields:
$0 = a + b - 3c$

This means, for example, that if $a = b = c = 1$, there can be no solution to the equations, since that would imply that $0 = -1$. On the other hand, if, say, $a = 0$, $b = 3$, and $c = 1$, then there is a simultaneous solution to the equations, but it isn't unique. One solution, for example, is given by $A = -1$, $B = 1$ and $C = 0$; another is given by $A = -3$, $B = 0$ and $C = 1$. That these are both solutions can be seen by substitution.

In general, whenever $a + b - 3c = 0$, one can choose C arbitrarily, then, based on the second equation, set $B = c - C$ and, based on the first equation, set $A = 2c – b -2C$. One checks that all three equations are satisfied, as follows:

$A + B + 3C = (2c – b -2C) + (c – C) + 3C = 3c – b = a$ [This last equality, because $a + b - 3c = 0$]
$-A + 2B = -(2c – b -2C) + 2(c – C) = b$
$B + C = (c – C) + C = c$

Regarding this system of equations the general situation has been understood

since the nineteenth century.[17] One calculates the so-called *determinant* of the linear system. (The determinant is a number that is calculated from a formula based on the coefficients of the set of equations.) If the determinant is non-zero, there is a unique solution. If the determinant is zero then, depending on the particular values of a, b, and c, there will either by no solutions or an infinite number of solutions.

But there is another way to look at the problem, one more closely related to the vector space perspective. In my last example, I deliberately chose vectors $\mathbf{v}_1$, $\mathbf{v}_2$, and $\mathbf{v}_3$ that were *linearly dependent*,[18] meaning, by definition, that there exist numbers A_1, A_2, and A_3, not all of them zero, such that

$A_1\mathbf{v}_1 + A_2\mathbf{v}_2 + A_3\mathbf{v}_3 = 0$
In this case,
$2\mathbf{v}_1 + \mathbf{v}_2\ \mathbf{v}_3 = 0$

This means, for example, that the last vector in this sum can be expressed as a *linear combination* of the previous two.[19] Specifically,

$\mathbf{v}_3 = 2\mathbf{v}_1 + \mathbf{v}_2$

This means, in turn, that any vector that can be expressed as a linear combination of the three vectors $\mathbf{v}_1$, $\mathbf{v}_2$, and $\mathbf{v}_3$, can already be expressed as a linear combination of the first two. Indeed, the sum $A\mathbf{v}_1 + B\mathbf{v}_2 + C\mathbf{v}_3$, by substitution of the expression for $\mathbf{v}_3$, yields:

$A\mathbf{v}_1 + B\mathbf{v}_2 + C\mathbf{v}_3 = A\mathbf{v}_1 + B\mathbf{v}_2 + C(2\mathbf{v}_1 + \mathbf{v}_2) = (A + 2C)\mathbf{v}_1 + (B + C)\mathbf{v}_2$
To put this point another way, any linear combination of the form $A\mathbf{v}_1 + B\mathbf{v}_2$ has the form:
$A\mathbf{v}_1 + B\mathbf{v}_2 = A(1, -1, 0) + B(1, 2, 1)$
$= (A, -A, 0) + (B, 2B, B) = (A + B, -A + 2B, B)$
And, by inspection, the coordinates (a, b, c) of any vector of this form are related by the equation:
$a + b – 3c = 0$

The three vectors cannot span R^3 because they are linearly dependent. To span R^3, to be serviceable as a basis for R^3, they must be linearly independent, the

opposite of being linearly dependent. To state this condition more positively, vectors $\mathbf{w}_1$, $\mathbf{w}_2$, …, $\mathbf{w}_n$ are linearly independent if and only if

$A_1\mathbf{w}_1 + A_2\mathbf{w}_2 + \ldots + A_n\mathbf{w}_n = 0$ implies $A_1 = A_2 = \ldots = A_n = 0$[20]

By the argument I made for the last example, vectors are linearly independent when it is impossible to express one of the vectors as a linear combination of the others and are linearly dependent exactly when it is possible to express one of the vectors as a linear combination of the other.

In the last example, I showed that $\mathbf{v}_3 = 2\mathbf{v}_1 + \mathbf{v}_2$, which is equivalent to $2\mathbf{v}_1 + \mathbf{v}_2$ $\mathbf{v}_3 = 0$. That is to say $\mathbf{v}_1$, $\mathbf{v}_2$, and $\mathbf{v}_3$ are linearly dependent. Conversely, there is clearly no way to express any of (1, 0, 0), (0, 1, 0), or (0, 0, 1) as a linear combination of the other two. These vectors are linearly independent because

$a(1, 0, 0) + b(0, 1, 0) + c(0, 0, 1) = (a, b, c)$ can only be zero if $a = b = c = 0$.

So, as I said earlier, these vectors, (1, 0, 0), (0, 1, 0), and (0, 0, 1), are a basis of R^3. Recall that a set of vectors, $\mathbf{v}_1$, $\mathbf{v}_2$, …, $\mathbf{v}_n$, in a vector space V, form a basis if and only if any vector in the vector space V can be uniquely expressed as a linear combination of $\mathbf{v}_1$, $\mathbf{v}_2$, …, $\mathbf{v}_n$.[21]

One should expect, first, that any four vectors in R^3 are linearly dependent and, conversely, that any basis of R^3 contains exactly three vectors. One might, perhaps, try to confirm this expectation by writing down the appropriate system of simultaneous linear equations and arguing from the properties of determinants. However, there is a standard argument,[22]applying to vector spaces generally, that makes this unnecessary.

First an important distinction: In some vector spaces such as R^n, there is a finite limit to the number of linearly independent vectors one can find. In general, if there exists a finite set of vectors in a vector space V that “span” the vector space then the vector space is said to be finite dimensional. To span the vector space means that every vector in the vector space can be expressed as a linear combination of vectors selected from that finite set of vectors.

R^n, for example, is finite dimensional. On the other hand, the domain of real-valued continuous functions defined on the interval from 0 to 1 is not finite

dimensional. As another infinitedimensional example, the domain of polynomials, without limitation on degree, is not finite dimensional.

Suppose that the vector space V is finite dimensional. Then there exists a finite set of linearly independent vectors $\mathbf{v}_1$, $\mathbf{v}_2$, …, $\mathbf{v}_n$ that span the vector space V. Now suppose there is another finite set of linearly independent vectors $\mathbf{w}_1$, $\mathbf{w}_2$, …, $\mathbf{w}_m$ that also spans the space. Here, *n* and *m* are positive integers. What can one conclude about the relationship of n (the number of v vectors) and m (the number of w vectors)? One should expect, and, indeed one finds, that m = n.

To see this, I follow an argument presented in Halmos.[23]To begin with, in any finite ordered set of linearly dependent non-zero vectors, there is a first vector that is a linear combination of all the preceding ones. Assuming that none of the vectors is zero, that vector cannot be the first vector. As one continues to consider additional vectors from the list one must ultimately come to a vector that first spoils the linear independence since the entire set is linearly dependent. Linear dependence means that there is a nontrivial linear combination of the vectors that sums to the zero vector. Non trivial means that at least one coefficient of the linear combination is non-zero. But, in particular, the coefficient of the most recently added vector cannot be zero, because otherwise the relationship of linear dependence would, contrary to assumption, hold without it. But that means one can solve the equation for the most recently added vector, simply by dividing the entire relationship by the coefficient of the last vector and then moving everything else to the other side of the equation.[24] This is essentially what I did in the example of linear dependence that I discussed earlier.

But how does this bear on my question?
Suppose, again, that sets of linearly independent vectors $\mathbf{v}_1$, $\mathbf{v}_2$, …, $\mathbf{v}_n$ and $\mathbf{w}_1$, $\mathbf{w}_2$, …, $\mathbf{w}_m$ both span V. Since the w vectors span V, then v_1 is a linear combination of the w vectors and, therefore, the set

$\mathbf{v}_1, \mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_m$

is linearly dependent. So one of the $\mathbf{w}_i$ vectors is a linear combination of the previous vectors in the list. Remove the first such vector. Since the removed vector is a linear combination of the previous vectors, the vectors that remain still span the entire vector space V.

Now add another v vector after the previous v vector. The result is the set
$\mathbf{v}_1, \mathbf{v}_2, \mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_m$

where one of the w vectors is missing. This set is linearly dependent because the set, without the added v vector, already spans the space, which means that the new vector is a linear combination of all the vectors in the previous set. So, once again, find the first vector that is a linear combination of the previous ones and discard it. The v vectors are all linearly independent so the discarded vector will have to be one of the w vectors. Continue in this way until all of the v vectors have been added to the list, discarding, each time, exactly one of the w vectors. At each step the list of vectors that remains spans the entire vector space V.

If one were to run out of w vectors before the end, the implication would be that a subset of the v vectors already spans the space. But this cannot be because the entire set of v vectors is linearly independent. So there must be at least one w vector remaining in the list as the last v vector is introduced. It follows that $n \leq m$. But one could also make the same argument in reverse, starting with the v vectors and adding w vectors one by one. But this time one would conclude that $m \leq n$. The two inequalities, taken together, imply that $m = n$.[25] In conclusion, if V is a finite dimensional vector space, any two maximal sets of linearly independent vectors contain the same number of vectors. That number is called the dimension of the vector space.

It follows that R^n is an n-dimensional vector space since one can exhibit a basis, consisting of n vectors, namely $\hat{e}_1, \hat{e}_2, \ldots, \hat{e}_n$, (as defined above for n = 3).

The argument that I gave earlier to show that R^3 is isomorphic to the vector space of quadratic polynomials extends to show that any two vector spaces of the same dimension n are isomorphic. My presentation in the earlier case was informal: I simply showed how sums of vectors and products of vectors by numbers lined up exactly between the two vector spaces. In effect, in providing a correspondence, I specified a map between them by associating (1, 0, 0) to x^2, (0, 1, 0) to x and (0, 0, 1) to 1 (the quadratic polynomial with zero coefficients for powers of x and 1 for the constant term.) I matched up basis vectors and, thereby, matched up all corresponding linear combinations of basis vectors. Implicitly my correspondence was a map from one vector space to the other. But I did not give that map a name.

To treat the general case, I will follow exactly the same procedure, but, in this case I give that map a name, namely T.
Begin by choosing bases for each vector space, say $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ for the vector space V and $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_n$ for the vector space W. Define the correspondence map T from V to W by setting first $T(\mathbf{v}_i) = \mathbf{w}_i$ for each integer i between 1 and n. Having, thus, lined up the basis, one next matches corresponding linear combinations of basis vectors. In this fashion, one defines T generally, as applying to any linear combination of basis vectors in V by the formula

$T(A_1\mathbf{v}_1 + A_2\mathbf{v}_2 + \ldots + A_n\mathbf{v}_n)$
$= A_1T(\mathbf{v}_1) + A_2T(\mathbf{v}_2) + \ldots + A_nT(\mathbf{v}_n)$

I write the formula in this way to emphasize the dependence of the mapping on the chosen values $T(\mathbf{v}_i)$ for i = 1, 2, … , n. But this amounts to, i.e., reduces to
$T(A_1\mathbf{v}_1 + A_2\mathbf{v}_2 + \ldots + A_n\mathbf{v}_n) = A_1\mathbf{w}_1 + A_2\mathbf{w}_2 + \ldots + A_n\mathbf{w}_n$

Thus the map T matches the linear combination $A_1\mathbf{v}_1 + A_2\mathbf{v}_2 + \ldots + A_n\mathbf{v}_n$ in V to the linear combination $A_1\mathbf{w}_1 + A_2\mathbf{w}_2 + \ldots + A_n\mathbf{w}_n$ in W, simply replacing each basis vector in V by the corresponding basis vector in W. This is exactly the process I followed in comparing quadratic polynomials with R^3.

Clearly this map T is reversible. If T maps each V basis vector to a corresponding basis vector in W, then the reverse, which is normally written T^{-1}, maps in the opposite direction, mapping each basis vector in W to the corresponding basis vector in V. Thus:

$T^{-1}(A_1\mathbf{w}_1 + A_2\mathbf{w}_2 + \ldots + A_n\mathbf{w}_n) = A_1\mathbf{v}_1 + A_2\mathbf{v}_2 + \ldots + A_n\mathbf{v}_n$

One notices the same substitution of basis vectors, but, this time, in the opposite direction. In sum, like the earlier example, T sets up an exact one-to-one correspondence between the two vector spaces V and W that preserves all vector space operations between the two spaces.

If, for example, I take a linear combination of two vectors in V, I can either perform the calculation in V or I can map the vectors to W, perform the calculation there, and then apply the inverse mapping. Because the mapping T, from a formal standpoint, is just a relabeling of basis vectors: Rename the V

basis vectors to $\mathbf{w}_1, \ldots, \mathbf{w}_n$, perform the operations, and then rename them again to $\mathbf{v}_1, \ldots, \mathbf{v}_n$.

In showing that any two finite dimensional vector spaces have the same structure, i.e., are isomorphic, I have shown, in particular, that any finite dimensional vector space is isomorphic to R^n for an appropriate value of n.

As a mathematician might put it, finite dimensional vector spaces are completely *classified* by their dimension, *classified*, one sometimes says, *up to isomorphism*. Qua vector spaces, there is no *structural* difference between, say, R^3, and quadratic polynomials. As different as they are in other respects, considered as vector spaces, such differences are omitted measurements.

The issue of classification arises precisely because some vector spaces have a different structure, *qua vector spaces*, than others. In this, vector spaces, as mathematical domains, differ from numbers: There is a unique mathematical domain of real numbers, but multiple mathematical domains consisting of vectors. In this respect, it is the number domain that is atypical in its uniqueness.[26]

Typically, there are multiple structurally distinct mathematical domains of a particular type. General speaking, for any general category of similar mathematical domains, there is a concept of isomorphism, a one-to-one mapping that preserves the entire constellation of structural relationships that characterize the scope of each domain. In the case of vector spaces these relationships, that an isomorphism must preserve, are addition of vectors and multiplication of vectors by numbers.

In the earlier example of a topological space, considered *only* as a topological space, an isomorphism is a one-to-one map that preserves open sets, the defining structural characteristic of a topological space, as described in Chapter 6. If f is an isomorphism between topological spaces X and Y, this means, first, that for any open subset U of X, its image f(U) is an open subset of Y and, conversely that if V is an open subset of Y and f^{-1} is the inverse map of f, then $f^{-1}(V)$ is an open subset of X. Here, by definition, if U is a subset of X, f(U) is the set consisting of all elements y in Y for which y = f(u) for some element u of U. Similarly, $f^{-1}(V)$ is the set consisting of all elements x in X for which $x = f^{-1}(v)$ for some element v of V.

Since the mapping f is one-to-one, the isomorphism establishes a one-to-one correspondence between the open sets, respectively of X and Y, the defining characteristics of their structures qua topological spaces.

A central problem with regard to any important type of domain is to classify mathematical domains of that type, to determine the structural ways in which two domains of that type can differ from each other, differ in the set of relationships that specifically characterize that type of domain. It is to find ways to measure the structural differences between two domains of a particular type. The case of finite dimensional vector spaces is one of the easiest cases. One number, the number of dimensions, is all that is required to identify a class of isomorphic vector spaces.

Choosing a basis of vectors in a finite dimensional vector space and then using them to compute is the bridge between the geometric perspective and the measurement perspective. In the earlier part of this section I started with a basis already at hand and proceeded to solve systems of simultaneous equations. That is the measurement perspective.

Later on, I presented Halmos's abstract argument that the number of vectors in a maximal set of linearly independent vectors characterizes the vector space: that this number is independent of which particular maximal set of vectors I might happen to come up with. Halmos's argument demonstrates the power of the abstract perspective: its ability to cover a wide terrain without getting bogged down in irrelevant details, and to show clearly the underlying principles involved. But, to the point of this discussion, the abstract perspective of that argument was the geometric perspective.

One can sum it up another way. A vector is a quantity and a basis is a set of quantities. Coordinates are numbers; coordinates are measurements. A focus on basis vectors is a geometric perspective; a focus on coordinates is the measurement perspective.

One can switch between the two perspectives because these perspectives are, in fact, two different perspectives of the very same unitary relationship, a relationship that is viewed from two different directions: It is the relationship, on the one hand, between the existents, or attributes, that one measures and, on the other hand, the measurements that one makes of those existents or attributes.

There are, as this section illustrates, advantages to both perspectives, to the

geometric perspective on quantities and to the measurement perspective. One needs both perspectives. And full understanding requires the integration of the two perspectives.

Matrices and Linear Transformations Matrices

The core *concept* of the modern theory of *linear algebra* is that of a vector space. But, as its core *concern*, linear algebra is about solving simultaneous linear equations. In this respect, linear algebra is far older than the modern theory of vector spaces; it is older, even, than algebra. As Kleiner points out,[27] the Babylonians, 4000 years ago, knew how to solve a system of two equations in two unknowns.

The need to solve simultaneous equations arises, in some form, in every branch of mathematics. When Euclid intersected two lines he was, from the perspective of analytic geometry, solving, geometrically, a system of two equations in two unknowns. It would be apt to say that all mathematical roads lead to linear algebra.

Although the physical concept of a *vector* dates back to Archimedes, the concept of a vector *space* came much later. In general, the key concepts in the modern theory of vector spaces made their debut *prior to* the formal definition of a *vector space*, by Peano, in 1888. Peano's definition was inspired by Grassman's earlier, little studied, 1844 work "whose aim was to construct a coordinate-free algebra on n-dimensional space."[28]

Among these central concepts of linear algebra are determinants, matrices, linear transformations, and linear independence. The concept of *determinants*, central to a systematic solution of systems of n simultaneous equations in n unknowns, was defined by Leibniz in 1693.[29] *Matrices*, following many precursors in the seventeenth and eighteenth centuries, were formally introduced by Cayley in 1850, having been given their modern English name by Sylvester in that same year. Cayley notes that matrices "comport themselves as single entities."[30] *Linear transformations*, a closely related concept, dates back to the seventeenth century.[31] Finally, Euler, in the eighteenth century, investigating linear differential equations, expressed the general solution as linear combinations of linearly independent solutions.[32]

All of these concepts relate, in some way, to systems of simultaneous equations. The discovery of the modern concept of vector spaces culminated an inductive process that spanned millennia. By the time the definition of *vector space* had been crystallized, much was already known about the essential characteristics and the diverse applications of vector spaces. The vector space concept organizes and systematizes all of these diverse strands. And it does so, fundamentally, by providing a geometric perspective.

But vector spaces provide that geometric perspective in a form that, thanks to analytic geometry, is amenable, as well, to analytic treatment, amenable to calculation. As such, for my purposes, the theory of vector paces provides a laboratory in which one can observe the interactions and interconnections between the geometric perspective and the measurement perspective in mathematics. This interconnection is nowhere more evident than in the relationship between matrices and linear transformations.

The simplest system of simultaneous equations has two equations and two unknowns. Take, for example, the system

$$2x + 3y = 16$$
$$x + 2y = 10$$

One easily solves this directly. The unique solution is given by x = 2 and y = 4. But my interest, here, does not center on solving this equation. Rather, it centers on alternative ways of capturing or expressing the essential relationships embodied in this set of equations.

I begin with the matrix representation. A matrix is a rectangular array of numbers and/or variables.[33]

One expresses the very same equations in the following form:

$$\begin{pmatrix} 2 & 3 \\ 1 & 2 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 2x + 3y \\ x + 2y \end{pmatrix} = \begin{pmatrix} 16 \\ 10 \end{pmatrix}$$

The square array on the left is a matrix consisting of the coefficients of x and y in the set of equations. As for the remaining single-column matrices, one should think of them as column vectors, simply a different way of representing vectors in R^2 than the coordinate-pair representation that I used earlier. Thus the equality on the right, thought of as equating two vectors, is another expression of the two simultaneous equations.

Next, think of the middle column vector as defining a set of calculations in

Next, think of the middle column vector as defining a set of calculations in relation to the matrix and vector on the left. Specifically, think of the implied matrix multiplication operation as involving two separate calculations, one calculation, respectively, for each row of the column vector shown between the two equal signs. The first row (2x + 3y) of the column vector is calculated by multiplying the first row of the square matrix by corresponding entries in the column vector next to it and adding the results. One places this sum (2x + 3y) in the first row of the resulting column vector. Next, one performs the corresponding operation for the second row of the square matrix and places the result in the second row of the resulting column vector.

The result is the column vector between the two equal signs. In sum, the expression on the left is a way of indicating the set of calculations captured in the middle column vector.

The final equality on the right says that corresponding entries in the two column vectors are equal and can be thought of as a reduction of the two simultaneous equations to one equation between two column vectors.

The entire exercise, so far, can be regarded, simply, as a way of organizing one's calculations. The leftmost matrix organizes the four coefficients into a convenient array, an array corresponding to their positions in the system of equations. The column matrix, as I said, is like a coordinate pair, and can be thought of as a column vector. The right hand equality, again, simply equates two column vectors. Once one has understood, as a process, the meaning of the so-called "matrix multiplication" on the left, one can omit the intermediate column vector and regard the remaining matrix multiplication on the left and its equality to the column vector on the right as providing an alternate expression of the system of simultaneous equations.

So what is a matrix? Visually, it is an array. But it's an array that captures a set of instructions, a set of calculations on a column vector. So one should think of a matix as an array that specifies a set of calculations on a column vector.

Now, suppose one multiplies each of the two column vectors on the right by the carefully chosen matrix

$\begin{pmatrix} 2 & -3 \\ -1 & 2 \end{pmatrix}$ Multiplying the first of these column vectors, following the process just discussed, one finds,

$$\begin{pmatrix} 2 & -3 \\ -1 & 2 \end{pmatrix}\begin{pmatrix} 2x+3y \\ x+2y \end{pmatrix} = \begin{pmatrix} 4x+6y-3x-6y \\ -2x-3y+2x+4y \end{pmatrix} = \begin{pmatrix} x \\ y \end{pmatrix}$$ Multiplying the second column vector, one finds:

$$\begin{pmatrix} 2 & -3 \\ -1 & 2 \end{pmatrix}\begin{pmatrix} 16 \\ 10 \end{pmatrix} = \begin{pmatrix} 32-30 \\ -16+20 \end{pmatrix} = \begin{pmatrix} 2 \\ 4 \end{pmatrix}$$ Applying the same recipe, the same set of calculations to equal quantities yields equal results. It follows that

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 2 \\ 4 \end{pmatrix}$$

Notice what happened here. Earlier, I multiplied the x-y column vector by the coefficient matrix and got a complicated expression. But later, when I multiplied the *result* of the first multiplication by the second matrix, the result of the second multiplication was the original column vector. In effect, as the composite of the two steps, I multiplied the x-y column vector by 1. And, in the process, of course, I solved the system of equations.

What happens if I generalize this matrix multiplication a bit? Suppose that I arrange two square 2 by 2 matrices next to each other. Then, suppose that one regards the matrix on the right as just a pair of column vectors, each one acted on independently by a matrix on its left. Each entry in the resulting matrix is a separate calculation, as follows:

$$\begin{pmatrix} 2 & -3 \\ -1 & 2 \end{pmatrix}\begin{pmatrix} 2 & 3 \\ 1 & 2 \end{pmatrix} = \begin{pmatrix} 2*2-3*1 & 2*3-3*2 \\ -1*2+2*1 & -1*3*2*2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

The matrix on the right is called the identity matrix and is usually designated by the capital letter *I*, the letter I standing for the word *identity*. To see why the name is justified, consider the product:

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x \\ y \end{pmatrix}$$ Now consider the expression

$$\begin{pmatrix} 2 & -3 \\ -1 & 2 \end{pmatrix}\begin{pmatrix} 2 & 3 \\ 1 & 2 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}$$

Having defined matrix multiplication one now has two choices. For example, one can multiply the column vector, in succession, by the two matrices, starting with the right-most matrix. Or, alternatively, one can first multiply the two matrices together in the fashion I have just indicated and apply the result to the

column vector. Either way (and this is true for all sequences of matrix multiplications) one gets the same result, regardless of which operation one performs first.

The identity matrix acts like the number 1 when it multiplies a vector and it also acts that way when it multiplies another matrix. Finally, the matrix that I multiplied the coefficient matrix by to get the identity matrix is called the inverse of the coefficient matrix. To save writing, and state the relationship in general, use the letter A to designate the coefficient matrix. In this notation, one writes the inverse matrix as A^{-1}. How to calculate A^{-1} is a standard topic in linear algebra textbooks and there is a formula, involving determinants, for the inverse of a matrix.

Not all square matrices have inverses, but when they do, one always has $A\ A^{-1} = A^{-1}A = I$. Here, there is nothing special about two dimensions. In n dimensions, the identity matrix is the matrix with all zeros except for the diagonal that runs from the upper left to the lower right. This diagonal, called the principal diagonal, contains the number 1 in every position. For example, the 3x3 identity matrix looks like this:

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Finally, to complete this introduction, or review, of matrices, one adds two matrices of the same shape by adding corresponding entries and one multiplies a matrix by a number by multiplying each entry in the matrix by that number. Considered only in regard to these two operations, the set of matrices of a particular shape can be regarded as a vector space. So everything one knows about vector spaces applies immediately to matrices. But, obviously, matrices are more than that. They have an independent wider interest, an interest, however, that derives from their relationship to a coordinatized vector space.

The vector space operations for matrices, the addition of two matrices and the multiplication of a matrix by a number, are meaningful because of the way that matrix multiplication applies to column vectors. If the sum of two matrices is applied to a column vector, the result is the same as multiplying the vector by each matrix separately and then adding the results. Multiplication of a matrix by numbers follows the same principle.

To illustrate the case of addition, assume that A and B are matrices and v is a column vector. Interpret Av to represent matrix multiplication of the column vector v by the matrix A from the left. Then Av is a column vector. If one interprets (A + B) as the matrix addition I described above, then (A + B)v = Av + Bv is a symbolic expression of the relationship between matrix addition and vector addition. One adds the matrices the way we do *because*, by doing so, one can add the matrices together instead of applying each matrix separately. The meaning of addition of matrices derives from the effect of matrix multiplication on vectors. The sum of two matrices represents the sum of the effects of the two matrices on vectors.

This continues a pattern that I have pointed out in this and earlier chapters. In general, addition in mathematics relates ultimately to addition of *numbers* in one way or another. We saw this for addition of fractions, addition of polynomials and functions, for addition of vectors, and, now, at one further remove, for additions of matrices. In general, two aspects enter into the development of derived notions of addition: first, the introduction of new units, as in the case of fractions, and, second, the multiplicity of distinct units within a single complex quantity. For functions, one adds values at every point; for vectors, one adds separate components.

To be a little more comprehensive, there is a final twist because some kinds of quantities have a cyclic character. For example, if one add an angle of 30^0 to itself a sufficient number of times, one ultimately reaches the sum of 360^0. And now one has a choice. If one is interested in the final position, one regards this final total as 0. But if one looks at this operation as successive rotations, and is specifically interested in the amount of the rotation, one regards this total as 360^0, a total that becomes even higher with the next rotation. Both viewpoints are valid within their appropriate contexts. And, either way, these successive additions are, properly, called addition.

Conversely, one should notice that, if A is a matrix that acts on vectors v and w, and if a and b are numbers, a kind of distributive law holds. Specifically,

$$A(av + bw) = aAv + bAw$$

Considered as an array of numbers, matrices constitute a system of measurements. Each entry in every matrix is a number and each entry functions as a number: Each number is a positive or negative measurement of some

magnitude, a magnitude that is one of a complex of quantities that, in some context, are related somehow to each other. Just as the *column vectors* on which they act reflect a choice of coordinate system, a *matrix expression* reflects the choices of coordinate systems for the column vectors involved, first for the x-y coordinate system in the vectors *to which it is applied* and, secondly for the column vectors *that result from* the matrix multiplication. By the time matrices enter the scene, these coordinate systems have already been chosen and the coordinates for the matrix simply reflect those choices. The ultimate physical or external meaning of a matrix derives from the meaning of the vectors to which it is applied and the meaning of the vectors that result from that application.

A mathematical domain of matrices is a system of measurements; matrices are used to compute. In regards to their application to vectors, a matrix is a unitary constellation of separate measurements treated as a single complex measurement.

Linear transformations

Linear transformations (also known as linear operators) are a certain kind of function or mapping from one vector space V to another vector space W. There is no restriction on either vector space participating in the relationship. There is no requirement that they have the same dimension; either one or both vector spaces might be infinite dimensional. Whatever the details, a mapping L from V, taking values in W is a *linear transformation* exactly when, for any two vectors v_1 and v_2 in V and any two numbers a_1 and a_2, the following relationship holds:

$L(a_1v_1 + a_2v_2) = a_1L(v_1) + a_2L(v_2)$.[34]
As a simple example, the mapping from V to V that multiplies every vector by 2 is linear. One sees this, as follows:
$L(a_1v_1 + a_2v_2) = 2(a_1v_1 + a_2v_2) = 2(a_1v_1) + 2(a_2v_2) = a_12v_1 + a_22v_2 = a_1L(v_1) + a_2L(v_2)$
The mapping $L(ax^2 + bx + c) = ax^2 + bx$ is linear, as is the mapping
$R(ax^2 + bx + c) = (a + b)x^2 + bx + c$
One sees the first, for the sum of two quadratic polynomials, by the calculation:

$L((ax^2 + bx + c) + (Ax^2 + Bx + C)) = L((a + A)x^2 + (b + B)x + (c + C)) = (a + A)x^2 + (b + B)x$

$= (ax^2 + bx) + (Ax^2 + Bx)$
$= L(ax^2 + bx + c) + L(Ax^2 + Bx + C)$

A similar exercise shows linearity with respect to multiplication of a polynomial by a number and similar calculations show that the map R is linear, as well.

The mapping, from quadratic polynomials to linear polynomials, that takes the calculus derivative of the quadratic polynomial is linear. This follows from the familiar facts that the derivative of a sum is the sum of the derivatives and the derivative of a function that has been multiplied by a constant is equal to that constant multiplied by the derivative of the function. If I give this linear transformation a name, namely D_x, one has, for example:

$D_x(ax^2 + bx + c) = 2ax + b$
One can also define an anti-derivative mapping Int_x, also linear, from quadratic polynomials to third-degree polynomials:
$Int_x(ax^2 + bx + c) = (a/3)x^3 + (b/2)x^2 + cx$

Linear transformations are aptly named, because they map lines to lines. To see this in R^2, consider the following parameterization of a line in R^2:

$$t \rightarrow \begin{pmatrix} 2+3t \\ 4-2t \end{pmatrix} = \begin{pmatrix} 2 \\ 4 \end{pmatrix} + t \begin{pmatrix} 3 \\ -2 \end{pmatrix}$$ Applying any linear operator to the right hand side, one finds:

$$L\left(\begin{pmatrix} 2 \\ 4 \end{pmatrix} + t\begin{pmatrix} 3 \\ -2 \end{pmatrix}\right) = L\left(\begin{pmatrix} 2 \\ 4 \end{pmatrix}\right) + t\, L\left(\begin{pmatrix} 3 \\ -2 \end{pmatrix}\right)$$

The expression on the right has the form of a parameterized line. This argument depends not at all on the nature of the particular vector space involved. Indeed, if **a** and **b** are fixed vectors in a vector space V, then **a** + t**b** is a parameterized line in V. Applying a linear transformation L, to this expression yields:

$L(\mathbf{a} + t\mathbf{b}) = L(\mathbf{a}) + tL(\mathbf{b})$
This is a parameterized line in the image vector space, say W, in which the fixed vectors if **a** and **b** in V are replaced by the fixed vectors L(**a**) and L(**b**) in the vector space W.

The action of any linear map on a finitedimensional vector space is completely

determined by its action on any basis. If $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ is a basis for the vector space V then any vector in V can be written in the form $A_1\mathbf{v}_1 + A_2\mathbf{v}_2 + \ldots + A_n\mathbf{v}_n$ for numbers A_i. If L is any linear transformation, then successive applications of the linearity condition yields:

$L(A_1\mathbf{v}_1 + A_2\mathbf{v}_2 + \ldots + A_n\mathbf{v}_n)$
$= A_1L(\mathbf{v}_1) + A_2L(\mathbf{v}_2) + \ldots + A_nL(\mathbf{v}_n)$

Assuming that the value of L is given on each basis vector, that $L(\mathbf{v}_i)$ is known for I = 1 . . . n, this expression determines the value of L for all vectors in V.

Assume, for example, that L is defined on the standard basis of column vectors in R^3, as follows:

$$L\left(\begin{pmatrix}1\\0\\0\end{pmatrix}\right) = \begin{pmatrix}2\\3\\4\end{pmatrix}, \quad L\left(\begin{pmatrix}0\\1\\0\end{pmatrix}\right) = \begin{pmatrix}20\\30\\40\end{pmatrix}$$

$$\text{and} \quad L\left(\begin{pmatrix}0\\0\\1\end{pmatrix}\right) = \begin{pmatrix}-2\\-3\\-4\end{pmatrix}$$

Applying L to a general vector

$$L\left(\begin{pmatrix}x\\y\\z\end{pmatrix}\right) = x\begin{pmatrix}2\\3\\4\end{pmatrix} + y\begin{pmatrix}20\\30\\40\end{pmatrix} + z\begin{pmatrix}-2\\-3\\-4\end{pmatrix}$$

$$= \begin{pmatrix}2x+20y-2z\\3x+30y-3z\\4x+40y-4z\end{pmatrix} = \begin{pmatrix}2 & 20 & -2\\3 & 30 & -3\\4 & 40 & -4\end{pmatrix}\begin{pmatrix}x\\y\\z\end{pmatrix}$$

In other words, L is represented, for this basis, by the matrix:

$$\begin{pmatrix}2 & 20 & -2\\3 & 30 & -3\\4 & 40 & -4\end{pmatrix}$$

A matrix, then, is a way of expressing a linear transformation.[35]
Every matrix acts as a linear transformation and every linear transformation, at least in the finite dimensional case, can be represented by a matrix. So what is the difference between the two concepts?

The difference is one of perspective. A matrix is an array of numbers; its application depends completely on the choice of basis of the vector space, i.e., its coordinate system. A domain of matrices is a system of measurements. A linear transformation, conversely, is conceived geometrically as having a fixed meaning independent of coordinate system. If a coordinate system changes, a matrix *representation* of the linear transformation will need to change with it, but the meaning of the linear transformation, its effect on vectors, remains the same.
The difference, specifically, is in the context of what is being regarded as fixed. A matrix is an array of *numbers*; if one chooses different bases for the vector spaces, changing the coordinates of the vectors, without changing the numbers in the matrix, the *calculations* with respect to the new coordinates will remain the same, but the coordinates of the result will now specify different vectors in the vector spaces and the matrix calculations will, therefore, specify a different mapping, a different linear transformation. As it applies to vector spaces, the meaning of the matrix, of the array of numbers, changes when the meaning of the coordinates to which it applies changes.
So a matrix is something that acts on coordinates whereas, a linear transformation is something done to *vectors*. If one changes coordinates, a linear transformation keeps its meaning, but the matrix that would be required to specify that linear transformation needs to change. To this point, if a matrix is *viewed* as a *representation* of a particular linear transformation, then when the coordinates change, the matrix changes as well, in order to represent the same *transformation* in the context of the new coordinates.
There is a kind of duality here that one encounters over and over in mathematics. The source of that duality is almost always the same. It’s built into the nature of measurement, as involving a relationship. A relationship is a unitary phenomenon, a unitary fact. But any relationship can be viewed from two alternate perspectives corresponding, respectively, to the two things being related: One relationship; two perspectives. The interplay between geometry and systems of measurement is but one instance of this phenomenon.

Multiplication

If T is a linear transformation from vector space U to vector space V and S is a linear transformation from V to vector space W, one defines the composite function ST from U to W as

$(ST)(u) = S(T(u))$[36]

Read this as: The action of the composite function ST on a vector u in U consists of first applying T to obtain a vector T(u) in V and then applying S to that vector T(u) to obtain a vector in W. The parentheses that I placed around ST emphasize that the composite transformation, thus defined, can be regarded as a single conceptual unit, as a single transformation. Henceforth, I shall omit those parentheses.

I say that ST is a linear transformation. To see that, pick any two vectors u_1 and u_2 In U and numbers a_1 and a_2. Applying definitions and using the linearity of both S and T:

$ST(a_1u_1 + a_2u_2) = S(T(a_1u_1 + a_2u_2))$
$= S(a_1T(u_1) + a_2T(u_2)) = a_1S(T(u_1)) + a_2S(T(u_2)) = a_1ST(u_1) + a_2ST(u_2)$

The final result, $ST(a_1u_1 + a_2u_2) = a_1ST(u_1) + a_2ST(u_2)$, is the defining characteristic of a linear transformation, that linear transformation being ST.

As an important special case, if vector space U = V = W, then S, T, and ST are a linear mappings from, say, U to itself.
As examples involving polynomials, consider the following composite linear transformations involving differentiation and the anti-differentiation operator Int defined earlier:
$D_{x2}(x^3 + 2x^2 + 5) = D_xD_x(x^3 + 2x^2 + 5) = D_x(3x^2 + 4x) = 6x + 4$

$Int_xD_x(x^3 + 2x^2 + 5) = Int_x(3x^2 + 4x) = x^3 + 2x^2$
$D_xInt_x(x^3 + 2x^2 + 5) = D_x(1/4\ x^4 + 2/3\ x^3 + 5x) = x^3 + 2x^2 + 5$

Notice that $D_xInt_x \neq Int_xD_x$ and also that $D_xInt_x = I$ where I is the identity mapping I(v) = v.
Now consider a matrix example. Suppose S and T are represented by matrices. Starting with T, take, for example,

$T = \begin{pmatrix} 2 & 3 \\ 1 & 2 \end{pmatrix}$ and $v = \begin{pmatrix} x \\ y \end{pmatrix}$ Then

$$Tv = \begin{pmatrix} 2 & 3 \\ 1 & 2 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 2x+3y \\ x+2y \end{pmatrix}$$ Now, suppose that

$$S = \begin{pmatrix} 4 & 1 \\ 1 & 0 \end{pmatrix} \text{ and } w = \begin{pmatrix} r \\ s \end{pmatrix}$$ So

$$Sw = \begin{pmatrix} 4 & 1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} r \\ s \end{pmatrix} = \begin{pmatrix} 4r+s \\ r \end{pmatrix}$$ Then, applying the definition of composition,

$$STv = S\begin{pmatrix} 2x+3y \\ x+2y \end{pmatrix} = \begin{pmatrix} 4(2x+3y)+x+2y \\ 2x+3y \end{pmatrix} = \begin{pmatrix} 9x+14y \\ 2x+3y \end{pmatrix}$$ But

$$\begin{pmatrix} 9x+14y \\ 2x+3y \end{pmatrix} = \begin{pmatrix} 9 & 14 \\ 2 & 3 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}$$ So ST is represented by the matrix

$$\begin{pmatrix} 9 & 14 \\ 2 & 3 \end{pmatrix}$$ On the other hand, matrix multiplication also yields:

$$\begin{pmatrix} 4 & 1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 2 & 3 \\ 1 & 2 \end{pmatrix} = \begin{pmatrix} 9 & 14 \\ 2 & 3 \end{pmatrix} = ST$$

So the product of the matrices representing S and T is the matrix representing ST. What I called matrix multiplication, in accordance with standard usage, turns out to be the same thing as composition of linear transformations. As a first observation, this confirms the claim I made when I introduced the product: that the multiplication of matrices derives from the effect that matrices have on vectors.

What about this multiplication?
I have already argued that the operations that mathematicians call addition are generally, if not always, derived from addition of numbers. That there might be multiple units, e.g., independent vectors, involving multiple dimensions of measurement, does not change the principle. It only means that more than one unit is involved in the constellation of measurements that one is further integrating into a single conceptual unit. Cyclic cases, like odd versus even, provide a more complex example. In the even/odd example, there are only two measurements, odd and even. One adds, for example, odd plus even and gets odd. Yet, behind the scenes, numbers are still involved; one is simply omitting measurements, retaining only whatever remains upon division by two.
In general, whenever a binary operation is called *addition* there is always a zero, every element in the system of measurements has an additive inverse, and

addition is commutative. To say that addition is commutative means that, for any elements a and b contained in the system of measurements, a + b = b + a.
Not so with multiplication. It is true that there is generally an *identity* element, like the *identity matrix*. An *identity* is an element, *I*, such that, for any A in the system of measurements AI = IA = A. But even in the case of matrices there are many elements without multiplicative inverses. And one does not require that an operation be commutative to be called a multiplication. Matrices and linear transformations exemplify this point, as well. I've already pointed out, for example, that $D_xInt_x \neq Int_xD_x$. Here's a matrix example:

$$\begin{pmatrix} 2 & 3 \\ 1 & 2 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 1 & 2 \end{pmatrix} = \begin{pmatrix} 5 & 6 \\ 3 & 4 \end{pmatrix}$$ But

$$\begin{pmatrix} 1 & 0 \\ 1 & 2 \end{pmatrix}\begin{pmatrix} 2 & 3 \\ 1 & 2 \end{pmatrix} = \begin{pmatrix} 2 & 3 \\ 4 & 7 \end{pmatrix}$$

Where matrices are concerned, it is the exception for multiplication to be commutative. So why the appellation? Why call it multiplication?

The analogy to arithmetic operations is, certainly, not as strong for multiplication of matrices and linear transformations as it is for addition of matrices, linear transformations, and vectors. And mathematicians could, certainly, call this operation something else.

But they don't. When mathematicians "multiply" things, they think of it as a kind of multiplication, all the time remembering the respects in which it differs from the numerical operation with the same name. And I don't think this is an accident. And I don't think it's the wrong thing to do.

To begin with, there are some relatively superficial points. First, whatever one calls it, the notation, writing elements next to each other to indicate the desired operation, for example, is the most compact and suggestive notation available for the relationships being expressed. And that includes singling out the identity element. Indicating, the inverse of an element A by A^{-1} is similarly, and rightly, done generally, to mean, essentially, to undo the effect of A on something. And when one speaks of multiplying (as opposed to forming a "product", which is still more general) the associative law always holds: If A, B, and C are elements in a system of measurements or a system of quantities, it is always the case when A, B, and C are multiplied together, that (AB)C = A(BC). This means that one

can first multiply A and B, obtaining their product and then combine that product with C or, conversely and to the same effect, multiply B and C together to find *their* product and then combine that result with A. The results will be the same.

In this, however, one must always keep track of the order; one cannot interchange two elements without possibly affecting the result: It would *not* be true, in general, that (AB)C = A(CB).

When does one use the term *multiplication* in mathematics? There are two kinds of typical cases. First, there are cases that really do derive from numbers. Multiplying two polynomials together to get another polynomial is such an example. The value of the product polynomial at every point is, simply, the product of the values of the two multiplied polynomials. In cases like this, multiplication is commutative because of the relationship to ordinary multiplication.

The other situation, to which one typically applies the term, *multiplication*, arises as the *composition* of two functions or transformations. Generally, though there are exceptions, the term *multiplication* is applied *only* when the functions involved map a domain onto itself.

Matrices, as we saw, are coordinate representations of linear transformations; their product represents a composition of linear transformations. But they are a counterexample to the general tendency. Namely, one *can* multiply matrices that relate different vector spaces of, e.g., different numbers of dimensions as long as the domain of each matrix contains the range of the previous matrix (on its right).

In the typical case, when the functions under consideration map a domain into itself, there is always an identity mapping, an element that maps every element in the domain to itself. And, for a 1 to 1 map, there is an inverse that is just the backwards map. Finally, if there is an addition being considered as part of the system of measurements, the multiplication and addition should interact in the right way: there is the kind of distributive law that I have observed in the case of matrices.

Considering the general need for conceptual integration, these are probably sufficient considerations to justify extending the term, *multiplication*, widening

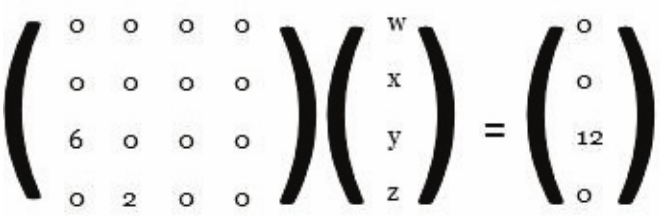
its meaning, by analogy, to a vastly broader context than its original meaning.

But there is also a deeper similarity.
In the realm of numbers, addition relates to *counting*, but multiplication relates more specifically to *measurement*, in the sense of *identifying the relationship to a unit*. It is because of multiplication that one cannot think of the number line as embodying an omitted unit, such as feet or decibels. One has to think of numbers on the number line as dimensionless, as *ratios*, as representing the *multiplicative relationship* between two magnitudes of the same kind, whatever that kind might be. A number, in the context of multiplication and measurement of continuous quantities, is a quantitative relationship to a unit.[37]
Therein lies the analogy to transformations: A transformation, or a function, represents a relationship between the elements that it relates. It is not always specifically a measurement qua relationship to a unit. But it does represent an *abstract* measurement, a quantitative relationship that can potentially represent part of an indirect measurement. In this sense, multiplication operations that represent compositions of transformations are closer to the essence of numerical multiplication than the multiplication of polynomials, based on their point-wise numerical values.
The role of abstract measurement was a major theme in the chapter on Euclid's Method, but this role has been part of the background of everything I have discussed since. And sometimes, in important cases, a transformation is more than an abstract measurement, a relationship between two quantities. Sometimes a transformation is a measure of symmetry. Measuring symmetry is the province of group theory, something that I will discuss in Chapter 8.

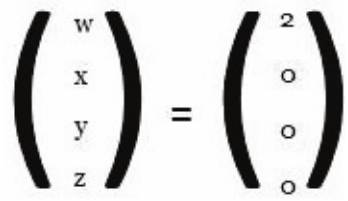
Kernels and Quotient Spaces

Linear Algebra, as I've said, is about solving simultaneous systems of equations. Looking at it abstractly, from a vector-space perspective, one wants to solve an equation $Av = w$, where v is in one vector space, V, w is, possibly, in another vector space, W, and A is a linear transformation, possibly represented, in a coordinate expression, by a matrix.

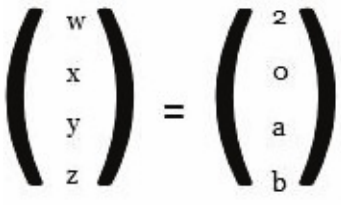
But solutions to simultaneous equations are not always unique. For example, consider the matrix equation:



This matrix equation represents four simultaneous equations, of which two are trivial. After discarding the two trivial equations, what remains are two equations: 6w = 12 and 2x = 0. Solving, one finds a particular solution:

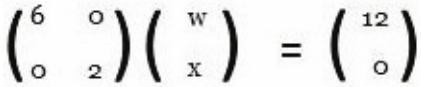


Are there any other solutions?
Notice that neither y nor z entered into the pair of equations. While the values of w and x were forced, the equations impose no restrictions whatever on either y or z. So the general solution is:

 for arbitrary numbers a and b.

And why is there no restriction on the values of y and z? Because y and z do not appear in the equations, 6w = 12 and 2x = 0. As far as this particular problem is concerned, y and z are totally irrelevant. The value of w is forced by the first equation and the second equation forces x to be zero. But y and z might as well not even be there. One could take them out of the matrix equation totally, and it wouldn't matter.

Indeed, the equations, 6w = 12 and 2x = 0, expressed as a matrix equation, amount to:



In this formulation, *of the same problem*, the variables y and z are nowhere in

sight. Nor are the variables, r, s, t, u, or v anywhere in sight. These latter variables, whatever they might be used to designate, have no more bearing on the actual problem, but, also, no less bearing, than the variables y and z. The only difference between the variables y and z, versus these other variables, is that y and z refer to some *specific* factors, initially *presumed* to be relevant, that, on further analysis, turn out to be irrelevant.

Had y and z been left out of the problem in the first place, there would have been a unique solution to the problem. But, because they are included, there is an entire family of solutions, namely,

$$\begin{pmatrix} w \\ x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 2 \\ o \\ a \\ b \end{pmatrix}$$

If one thinks about this geometrically, as a problem in a fourdimensional vector space, one finds that, in this example, four dimensions are included in the *statement* of the problem, but only two dimensions are actually *relevant* to the problem. So the fourdimensional problem reduces to a two dimensional problem.

Now consider a more numerically complicated example. Consider the equation

$$\begin{pmatrix} 1 & -1 & o \\ 1 & 2 & 1 \\ 3 & o & 1 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 3 \\ 5 \end{pmatrix}$$

Without showing the derivation, I note, and one can check, that x = 2, y = 1, z = -1 is a particular solution to this equation. For, by direct computation:

$$\begin{pmatrix} 1 & -1 & o \\ 1 & 2 & 1 \\ 3 & o & 1 \end{pmatrix}\begin{pmatrix} 2 \\ 1 \\ -1 \end{pmatrix} = \begin{pmatrix} 1 \\ 3 \\ 5 \end{pmatrix}$$

I now ask the same question as I did for the first example: Is the solution x = 2, y = 1, z = -1 unique? There is a general conceptual approach to such problems.

In the general formulation, Av = w, where v is a vector in a vector space V and w is a vector in a vector space W. As a matter of terminology, if w is chosen to

be 0 (the *vector* 0 in W) then the resulting equation Av = 0 is known as the *homogeneous* equation. Otherwise, when w ≠ 0, the equation Av = w is known as the *inhomogeneous* equation.

I'm interested in uniqueness. Accordingly, suppose that v_1 and v_2 are two solutions of the inhomogeneous equation. That is, suppose that v_1 and v_2 satisfy, respectively, the equations $Av_1 = w$ and $Av_2 = w$. It follows that

$Av_1 - Av_2 = w - w = 0$
By the linearity of A, $A(v_1 - v_2) = Av_1 - Av_2$ and, therefore
$A(v_1 - v_2) = 0$

So the *difference* $(v_1 - v_2)$ between two solutions of the *inhomogeneous* equation (Av = w) is a solution of the *homogeneous* equation, namely, Av = 0.

The set *of solutions* to the homogeneous equation Av = 0 has a special name. It's called the *kernel* of the transformation A.
My first example had a very simple kernel consisting of vectors of the form

$$\begin{pmatrix} w \\ x \\ y \\ z \end{pmatrix} = \begin{pmatrix} o \\ o \\ a \\ b \end{pmatrix}$$

For, clearly, the equation,

$$\begin{pmatrix} o & o & o & o \\ o & o & o & o \\ 6 & o & o & o \\ o & 2 & o & o \end{pmatrix}\begin{pmatrix} w \\ x \\ y \\ z \end{pmatrix} = \begin{pmatrix} o \\ o \\ 6w \\ 2x \end{pmatrix} = 0$$

holds precisely when w = x = 0.
By the nature of this example:

● Any two solutions to the *inhomogeneous* equation differ only in their last two coordinates, since the values of the first two coordinates of the solution were forced.

● The values of the variables a and b can be chosen freely. ● By the same token,

for the *homogeneous* equation, w and x must be zero, and, once again, the values of the variables a and b can be chosen freely.
● These solutions of the homogeneous equation form a vector space, namely the vector space consisting of vectors for which both w and x are zero.

In general, the solutions to a homogeneous equation, Av = 0, always form a vector space. To see this, start with the obvious fact that A0 = 0, where the first 0 is, again, the zero vector in V. Secondly, if v_1 and v_2 are solutions to the homogeneous equation, and a_1 and a_2 are numbers, then

$$A(a_1v_1 + a_2v_2) = a_1A(v_1) + a_2A(v_2) = 0 + 0 = 0$$

In other words, any linear combination of solutions to the homogeneous equation is, itself, a solution to the homogeneous equation. But this, closure under addition of vectors and multiplication by numbers, is the defining property of a subspace. So the solutions to the homogeneous equation form a subspace of the larger vector space to which they are already known to belong. And this means that the solutions to the homogeneous equation are, themselves, a vector space.

But how does this bear on the inhomogeneous equation? Simply this: If K is the kernel of a linear transformation A, and v_0 is a particular solution to the inhomogeneous equation, then the complete set of solutions is completely characterized as the set of vectors of the form $v_0 + v_k$ where v_k is contained in the kernel of A. First, as we have just seen, any two solutions differ by a solution to the homogeneous equation, in short, by a vector in the kernel. Conversely, one computes:

$$A(v_0 + v_k) = A(v_0) + A(v_k) = w + 0 = w$$

With this as a background, I return to our problem with the 3 X 3 matrix:

$$\begin{pmatrix} 1 & -1 & 0 \\ 1 & 2 & 1 \\ 3 & 0 & 1 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 3 \\ 5 \end{pmatrix}$$

for which x = 2, y = 1, z = -1 is a particular solution to this equation.

What is the kernel? In light of this discussion, one reformulates this question, as follows: What are the values of x, y, and z such that the following holds?

$$\begin{pmatrix} 1 & -1 & 0 \\ 1 & 2 & 1 \\ 3 & 0 & 1 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$ This matrix equation amounts to three simultaneous equations that can be solved by inspection:

$x - y = 0$
$x + 2y + z = 0$
$3x + z = 0$

My plan of attack for this particular example is straightforward: First *choose* a value for x. Next, use the third equation to force the value of z and use the first equation to force y. Finally, one checks that these values of x, y, and z also satisfy equation 2.

For example, if $x = 1$, then from the third equation, $z = -3$. And, from the first equation, one must have $y = 1$. One checks that these three values also satisfy the second equation.

In this example, one has a free choice for the value of x. But, once x is chosen, the corresponding values of y and z are forced.

Obviously, and as we've seen, any multiple of a solution to a homogeneous equation is also a solution. So, If a is any constant, it follows that all vectors of the form

$$a\begin{pmatrix} 1 \\ 1 \\ -3 \end{pmatrix} = \begin{pmatrix} a \\ a \\ -3a \end{pmatrix}$$

satisfy the homogeneous equation. On the other hand, we have already seen that $x = a$ implies that $y = a$ and $z = -3a$. In light of our general discussion, therefore, the general solution of the *inhomogeneous* equation is

$$\begin{pmatrix} 2 \\ 1 \\ -1 \end{pmatrix} + a\begin{pmatrix} 1 \\ 1 \\ -3 \end{pmatrix} = v_1 + av_k$$

where one sets $v_1 = \begin{pmatrix} 2 \\ 1 \\ -1 \end{pmatrix}$ and $v_k = \begin{pmatrix} 1 \\ 1 \\ -3 \end{pmatrix}$

Here, *a* is any constant number.

This second example involved more computation than the first and the answer is also more complicated. In the first example, the irrelevance of the last two variables was transparently obvious. There is no such transparency in this example.

Yet the same principle applies. Assuming that one's *sole* concern is finding a solution, there is an abundance of riches here. There is a one-parameter family of solutions; yet one requires but one solution. Differences, of the form av_K, don't matter; they do not affect the result. If one adds a multiple of v_K to a solution to the equation, the result is another solution to the equation. If one adds a multiple of $v_K$to a vector that is not a solution, the resulting sum is still not a solution. Either way, differences in the direction of v_K are irrelevant; they do not affect the outcome.

The contrast between this example and the first example is not a geometric contrast; rather, it involves a difference in one's choice of coordinates. Had I chosen a basis of R^3 for which the vector v_k were one of these *basis vectors*, the second example would have looked very much like the first.

To speak more generally, consider any matrix equation Av = w in relation to the kernel of the homogeneous equation Av = 0. To the extent one's *only* purpose is to solve Av = w, it doesn't matter which solution one chooses. The choices are *identical* with respect to their *image under the linear transformation A*. For this purpose, any two such choices are *equivalent*. Any two vectors in V that differ by an element k in the kernel of A will map to the *same* vector in W.

If u and v are vectors in V for which $u - v = v_K$, a vector in the kernel K, then

$u = v + v_K$

and

$Au = a(v + v_K) = Av + Av_K = Av$

Insofar as one's only concern is the value of Av, the two vectors u and v are equivalent. In effect, the difference between u and v can be ignored, *can be treated as an omitted measurement*.

Mathematicians call this sort of relationship an *equivalence relation*: Two vectors are regarded as equivalent, in this context, if and only if their difference is in the kernel K of the linear transformation A.

In general, an equivalence class is a set of vectors in a vector space that are distinguished as being equivalent from a particular perspective. In this respect, two vectors are equivalent by virtue of the fact that their difference lies in the specific subspace K.

These equivalence classes can be viewed as a vector space because it turns out, from the perspective of this notion of equivalence, the results, the equivalence classes, of linear combinations of vectors does not depend on which vectors are chosen, within their respective classes.

This is a difficult concept. I will complete the general discussion shortly. But the best way to visualize what is happening is to start with a simpler example.

Consider, then, vectors in R^3. For certain purposes, one, in effect, projects vectors into the xy plane. One regards two vectors as equivalent if they differ only in their zcomponent. Two vectors are treated as equivalent if their difference is a vector along the z axis, a vector that, expressed in row coordinates, takes the form (0, 0, z). One focuses on the x and the y coordinates, treats them in the usual way, and simply ignores the zcomponent as an omitted measurement, as irrelevant, on the usual principle: The zcomponent must have some value, but it may have any value.

If one takes a linear combination of two vectors in R^3, one can *continue* to restrict one's focus on the x and y coordinates because the values of the z coordinates of the two added vectors does not affect the x and y coordinates of the result.

In treating these vectors as equivalent, one focuses on the measurements that are *relevant in a particular context.*
When physicists restrict their attention to two dimensions, in analyzing a particular problem, this is exactly what they are doing: focusing on the

measurements that are relevant to a particular analysis and ignoring the rest. For example, in analyzing the effects of gravitation on a falling object, one typically studies the height z of the falling object as a function of time t. One ignores x and y. One ignores, for example, the effect of a breeze blowing in a north-eastern direction because one is not interested in the north-easterly drift. One can ignore this drift precisely insofar as it does not affect the height z, as a function of time.

To recap and continue the general discussion, if V is any vector space and K is a subspace of V, then vectors are considered equivalent, with respect to K, if their difference lies in the subspace K. The vector space that ignores distinctions among vectors occupying the same equivalence class, is called the *quotient space of equivalence classes with respect to K*. In standard notation, this vector space is written *V/K*.

A vector in V/K is simply a *vector in V* for which one omits certain distinctions, in which one treats any two vectors differing by a vector in K as the *same* vector. I have yet to show that, in fact, one can successfully treat V/K as a vector space. As a first step in that direction, I will specify the recipes for adding two equivalence classes and for multiplying an equivalence class by a number. Then I will illustrate how this recipe applies to a concrete example. Finally, I will provide a general argument to show that these operations are welldefined, that choices of particular representatives of equivalence classes do not affect the result.

This last point is the critical one: If it's truly the case that the distinctions between equivalent vectors don't matter, then the distinctions between their sums, for example, shouldn't matter either.

I start with the recipe. First, addition of vectors in V/K consists in;

1. choosing a representative from each equivalence class, 2. adding these representatives,
3. assigning the resulting sum to the *equivalence class* of the result.

Multiplication by a number is similar. One

1. chooses a representative from an equivalence class,
2. multiplies it by the number,
3. assigns the result to its equivalence class.

My claim, still in question, is that the equivalence classes of these results are unaffected by one's optional choices along the way. Let's see how this applies to my projection example. Suppose that one starts with a vector (x, y, z) for which

my projection example. Suppose that one starts with a vector (x, y, z) for which only the x and y coordinates are important. Let's say the x and y coordinates are (x, y) = (1, 2). Now take a second vector for which the x and y coordinates are (x, y) = (3, -1).

Let's choose representatives from their equivalence classes. One knows the first and second coordinates of the first vector, but the z coordinate could be anything. So, for the sake of the argument, choose the vector (1, 2, 4) as a representative of the first equivalence class. This vector is a representative because it has the correct x and y coordinates.

In similar fashion, for the sake of the argument, choose the vector (3,-1, 15) as a representative of the second equivalence class.
Their sum in R^3 is (4, 1, 19). The equivalence class of this sum consists of all vectors in R^3 for which the x coordinate is 4 and the y coordinate is 1.
There are two important points about this sum. First, the z component, 19 is irrelevant because it does not affect the equivalence class of the result. But, equally important, neither of the z components, 4 and 15, of the two *summands* affected anything but the z component of the *result*. Neither the x component nor the y component of the result was affected by these choices.
In effect, one might as well have left out the z component altogether, and written (1, 2) + (3, -1) = (4, 1).
The same point applies to multiplication by a number. Suppose, for example, one multiplies the first vector by 3. Choosing the same representative of the first vector that I chose earlier, one finds, 3 × (1, 2, 4) = (3, 6, 12). The equivalence class of this result consists of all vectors in R^3 for which the x coordinate is 3 and the y coordinate is 6.
Once again, the choice of z component, namely 4, for the vector (1, 2, 4), does not affect the x and y coordinates of the result and the z coordinate of the result does not affect the equivalence class of the result.
At least in this case, one can treat these equivalence classes as welldefined vectors, because the effect of vector addition and multiplication by numbers, the equivalence classes of the results, does not depend on which equivalent vector one chooses to subject to these operations. The results corresponding to different equivalent choices will always be equivalent.
Now the general situation is not so straightforward. The same principles apply, but their operation is not as transparent.
To make that general argument: Take two representatives u and v of equivalence classes u^ and v^ in V/K. Here, I'm simply using the ^ as a way to distinguish a

vector in V from its equivalence class in V/K. The vectors u and v are vectors in V; the equivalence classes u^ and v^ are, as I intend to validate, welldefined vectors in V/K. The vector u is a member of the set u^ and the vector v is a member of the set v^.

I need to show that the *equivalence class* of the sum does not depend on one's choices of the representative vectors u and v. One needs to show that the *equivalence class* of u + v, which I would like to write (u + v)^, is independent of my choices of u and v.

So assume that u_1 is equivalent to u and that v_1 is equivalent to v. I need to show that $u_1 + v_1$ is equivalent to u + v.

Since u_1 and u, are equivalent, their difference $u_1 - u$ is equal to some vector in K. Call this vector u_{1k}. That is, $u_1 - u = u_{1K}$, is contained in K. Likewise suppose that $v_1 - v$ is equal to some vector v_{1K} contained in K.

I need to show that the sum $(u_1 + v_1)$ is equivalent to the sum (u + v). But, by straightforward calculation, one finds $(u_1 + v_1) - (u + v) = (u_1 - u) + (v_1 - v) = u_{1K} + v_{1K}$. This sum is contained in K because K is a vector space. If I add two vectors in K, the result is in K.

A similar, but easier, argument applies to multiplication of u by a constant a. One needs to show that au_1 is equivalent to au. But this also follows by a calculation, namely: $au_1 - au = a(u_1 - u) = au_{1K}$. This product, au_{1K}, is a vector in K because, once again, K is a vector space.

Now how, from a reality-based perspective, should one look at this?

Notice that the elements of this vector space are, literally, in the conventional view, sets of vectors, sets of equivalent vectors. But these sets are condensed to a single unit, just as one does when one forms a concept. Mathematically, the equivalence class is a set, but it is a set viewed from a particular *conceptual* perspective. In this perspective, one is doing more than simply isolating vectors into disjoint *sets*, one is regarding any two vectors residing in the same set as *conceptually equivalent* in regards to a particular context.

One, as an incidental matter, can regard these vectors as elements in various sets. But one's interest in the vectors is *not* specifically as elements of the set. Rather, one is interested in them because they are *equivalent from a particular perspective*. One's use of set theory is helpful technically. But it is not the use of set theory that makes it meaningful. An equivalence class is a mathematical abstraction that transcends a set-theoretic perspective on mathematics.

Within a particular context, a particular quotient space should be regarded as a concept and as a mathematical domain. However, the resulting particular

quotient space is not a *permanent* conceptual unit that one would normally transfer to another context.

Nonetheless, *within the scope of a particular analysis*, it *is* a concept. The elements of each subset have been conceptually isolated according to a specific characteristic, such as their image under the mapping A. The elements of the set are the immediate referents of that conceptual distinction and they are also the link to their ultimate referents in the world, to all potential referents of the vectors in V itself. And the formation of that concept follows the fundamental principle of concept formation, the principle identified by Ayn Rand, of omitting measurements.

Recall my discussion of odd versus even numbers versus the classification of numbers by their remainder on division by 5 in Chapter 6. One forms a *concept* of even versus odd; it is a concept that children learn. But one does not form a permanent concept, a specific conceptual unit, to each possible remainder upon division by 5. One could not reasonably do so for each possible divisor.

Yet, in certain limited *contexts*, one does exactly that. So mathematicians need a *general* way of dealing with contexts in which a remainder is all that one cares about. And that is what they've done: One says $x = y \pmod 5$ precisely when $(x – y)$ is divisible by 5 (x and y being integers). And this is a general method that can apply to any divisor: One *also* says that $x = y \pmod 7$ precisely when $(x – y)$ is divisible by 7.

In viewing remainders in this way, one follows the same principle that one follows in forming quotient spaces of vector spaces. As I remarked at the beginning of Chapter 3, many of the most profound conceptions in mathematics make their first appearance in ordinary arithmetic.

This notion of a *quotient* is a very broad concept in mathematics that extends to many other systems of measurement and systems of quantities in mathematics. The formation of a *quotient*, of some sort, always follows the principle of omitting measurements and it is always done for the same reason – to restrict one's focus, in a particular context, to the relevant measurements in that particular context.

We will meet this concept of a quotient again, in the next chapter, for a very different mathematical domain.

Distances and Angles in Vector Spaces Distance Formula in Vector Spaces

According to the Pythagorean Theorem, the square of the distance from the origin (0, 0) of R^2 to a point $X = (x_1, x_2)$ is $x_1^2 + x_2^2$.[38] As an expression of this relationship, it is convenient to introduce the notations $|X| = \sqrt{(x_1^2 + x_2^2)}$ and, more importantly, $|X|^2 = x_1^2 + x_2^2$ as the square of the distance from (0,0) to (x_1, x_2).

Consider how this works in R^3. The square of the distance from (0, 0, 0) to $(x_1, x_2, 0)$ is, clearly, $x_1^2 + x_2^2$ because adding a coordinate simply adds an additional measurement, one that doesn't impact my measurements in the xy plane. Now consider the line from $(x_1, x_2, 0)$ to (x_1, x_2, x_3). This line is perpendicular (orthogonal) to the line from (0, 0, 0) to $(x_1, x_2, 0)$. The line from (0, 0, 0) to (x_1, x_2, x_3) represents the hypotenuse of a right triangle. The square of the distance from (0, 0, 0) to $(x_1, x_2, 0)$ having been determined as $x_1^2 + x_2^2$, a final application of the Pythagorean Theorem yields, for vectors in R^3:

$|X|^2 = x_1^2 + x_2^2 + x_3^2$

Refer to Figure 10 regarding this argument:

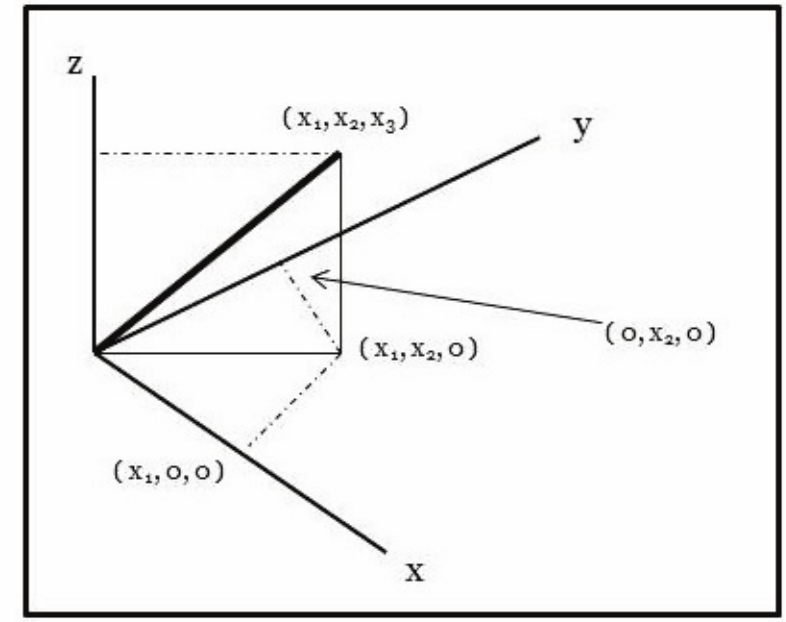


Figure 10

In this instance, although every step in the argument focused on a distance within a plane, the symbol $|X|^2$ pertains to vectors in R^3 instead of R^2. A less ambiguous notation would distinguish these two applications of the symbol $|X|^2$ by adding a subscript: thus, $|X|_{22}$ versus $|X|_{32}$.

What about R^4? Simple analogy would suggest:
$|X|_{42} = x_{12} + x_{22} + x_{32} + x_{42}$

An analogy is certainly needed, but it is needed at an earlier point in the conversation. Notice that in going from R^2 to R^3, one applied, at each point, a two-dimensional analysis. In the first step, one looks at the xy plane for which $x_3 = 0$. Then having established $|X|_{22}$ for the vector from (0, 0, 0) to (x_1, x_2, 0), one introduces a third point, namely (x_1, x_2, x_3). These three points determine a plane. And, in that plane, the line from (x_1, x_2, 0) to (x_1, x_2, x_3) is perpendicular to the line from (0, 0, 0) to (x_1, x_2, 0). So, the Pythagorean Theorem applies to the triangle determined by the vertices (0, 0, 0), (x_1, x_2, 0), and (x_1, x_2, x_3). What makes the argument work is the fact that the direction of the new axis is orthogonal to the other two.

It is at this point that one is forced to make an analogy. With each new axis, one *thinks* of that additional axis as being orthogonal to all of the previous axes.

And even that involves something a little deeper: a kind of homogeneity assumption.

For consider the difference between three planes in R^4: Consider the difference between the plane consisting of points (x_1, x_2, 0, 0), the plane consisting of points (0, 0, x_3, x_4), and the plane consisting of points (0, x_2, x_3, 0). Prior to the addition of the fourth axis, one could regard the first and the last of these as Euclidean planes, subject to the Pythagorean Theorem. However, the second of these planes can only be regarded as Euclidean if one thinks of the fourth axis as being orthogonal to the other three.

But if one does not think of it this way, then one is saying that there is something different about this fourth axis or something special about the first three axes. If there is value in the measurement that one applies to the first three dimensions then it is arbitrary to withhold it from the fourth.

Geometrically, the issue is this: In any mathematical analysis involving four independent variables, if one regards those four variables as being similar in kind, as being similar in the kinds of magnitudes one represents, then *none* of those dimensions is really a spatial dimension, because space is three dimensional. (Here, contrary to much of this chapter, I refer to geometry in the spatial sense.)

Notwithstanding, there is a context for which all four coordinates *might* all

represent spatial coordinates, namely a problem involving multiple bodies, each measured spatially. In such cases, it is not at all unreasonable to regard the coordinates of the second body to be orthogonal to the coordinates of the first. Although this represents an extension of the concept *orthogonal*, that extension is a natural one.

But the analogy is broader than such cases. And, in general, the resort to analogy does not occur when one adds the fourth axis to the first three. It occurs when one adds the second axis, regards the second axis as comparable to the first, and regards the new axis as orthogonal to the first. The essential analogy consists in applying geometric (spatial) measures of distance and angles to nongeometric settings.

Why might one apply concepts of distance and angles to nongeometric settings? To measure an approximation: A distance function provides a way to measure the difference between two constellations of quantities by a single number. Consider, first, the case of magnitudes. For magnitudes, one measures the difference between individual magnitudes in numerical terms, based upon a choice of standard.

But, where a multitude of different dimensions are concerned, one cannot study limiting processes without some way to ascertain that two *constellations* of measurements are close to each other. One needs some category or categories of measurements to make this determination.

As I discussed in Chapter Six, one can define topologies to achieve this goal, in general, without defining a distance function embracing all of the relevant dimensions into one formula. So defining a metric, specifying a measurement of distance to integrate the separate differences along each axis into a single numerical measurement, is not a strict necessity.

Still, there is often value in finding a single number to measure and summarize a difference. This is especially true when the various axes are commensurate in some way. But distancemeasures are valuable more generally.

Take a case of two independent variables involving totally different kinds of quantities, say weight and volume. In general, measurements of weight and volume cannot meaningfully be compared. Nonetheless, the *importance* of differences in two variables sometimes *can* be compared and even a very rough comparison is sometimes helpful. In some contexts, one can assess how a difference, in pounds, for one variable, *compares in importance* to a difference, in cubic feet, for the second variable. For example, based on limitations in one's ability to make approximations, one might weight an error of ¼ of a pound to be roughly equal in importance to an error of 1/8 of an inch. In a broader numerical example, one might assign weighting factors to signify, for example, that a

numerical difference of 2 in the first variable is roughly equal, in importance, to a numerical difference of 5 in the second variable.
Once such weights are given, one can adjust the units for the two dimensions to make them homogeneous: In these adjusted units, a difference of 1 in the x direction has roughly the *same importance* as a difference of 1 in the y direction. Treating these independent variables as orthogonal, one applies the distance formula to estimate, with one number, the importance of any difference between two *pairs* of values for the two variables.
This procedure should be regarded as a concept of method, as a way of meaningfully capturing, of measuring, certain relationships and differences that do exist and are important in some way.
With all that said, let us grant a way of looking at R^n that treats every axis as orthogonal to the other axes and, for some legitimate purpose, treats the magnitudes in each direction as comparable. With that as a starting point, the extension of the Pythagorean Theorem to R^n follows from induction.
For, suppose that one has already established that

$$|X|_{n-1\ 2} = x_{12} + x_{22} + \ldots + x_{n-12}$$

In R^n, this is, then, the distance from $(0, 0, \ldots, 0)$ to $(x_1, x_2, \ldots + x_{n-1}, 0)$. The line from $(x_1, x_2, \ldots + x_{n-1}, 0)$ to $(x_1, x_2, \ldots + x_{n-1}, x_n)$ makes a right angle with the first one. Therefore, by the Pythagorean Theorem,

$$|X|_{n\ 2} = |X|_{n-1\ 2} + x_{n2} = x_{12} + x_{22} + \ldots + x_{n2}$$

Inner Products: Measurement of Angles

With this definition, one has a measurement of distance. What about angles? The best place to start is with the two dimensional case. Consider the diagram:

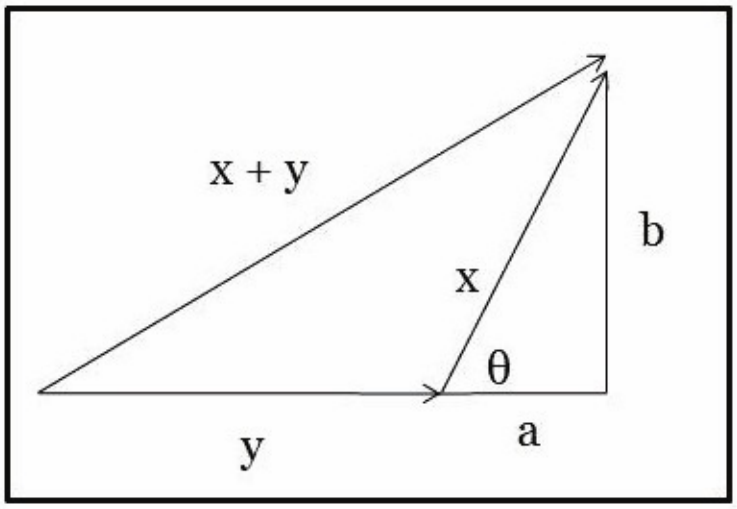


Figure 11

x and y are vectors, x + y is their vector sum. To avoid ambiguity I use the expressions |x|, |y|, and |x + y| to designate their lengths. The angle between the vectors x and y is θ. And the cosine of that angle is given by $\cos\theta = a/|x|$ or, in the form in which I intend to apply the formula, $|x|\cos\theta = a$.

By the Pythagorean Theorem,
$|x|^2 = a^2 + b^2$
and
$(|y| + a)^2 + b^2 = |x + y|^2$
Expanding the left hand side of the second equation and substituting for b^2 from the first equation, one obtains:
$|y|^2 + 2a|y| + a^2 + |x|^2 - a^2 = |x + y|^2$
Cancelling terms and moving two terms to the right hand side of the equation yields:
$2a|y| = |x + y|^2 - |x|^2 |y|^2$

As a matter of historical interest, Euclid establishes this very relationship in Book II, Proposition 12.[39]
The term on the right hand side is defined entirely in terms of the lengths of the three vectors x, y, and x + y. The left hand side, substituting for *a* from the cosine formula, becomes $2|x||y|\cos\theta$. With this substitution, the formula becomes:
$2|x||y|\cos\theta = |x + y|^2 - |x|^2 - |y|^2$

Solving for cosθ,
$\cos\theta = (|x + y|^2 - |x|^2 - |y|^2)/ 2|x||y|$

The expression for cosθ is defined entirely in terms of the lengths of the various vectors.
It is worth looking at the numerator $|x + y|^2 - |x|^2 - |y|^2$ in coordinates. One has,

$|x + y|^2 - |x|^2 - |y|^2$
$= ((x_1 + y_1)^2 + (x_2 + y_2)^2) - (x_{12} + x_{22}) - (y_{12} + y_{22}) = 2x_1y_1 + 2x_2y_2 = 2(x_1y_1 + x_2y_2)$

If one reflects on this calculation, it is immediately apparent that this calculation, if carried out in R^n, would result in the expression

$|x + y|^2 - |x|^2 - |y|^2 = 2(x_1y_1 + x_2y_2 + \ldots + x_ny_n)$

With this motivation, given the relationship of this expressiontocosθandconsideringitsrelationshiptothelengthsof the vectors involved, define the expression $< x, y >$ by the following formula:

$< x, y > = ½(|x + y|^2 - |x|^2 - |y|^2)$

Notice that $< x, x > = |x|^2$. This follows, upon substituting x for y in the formula for $< x, y >$.
As one should notice from the coordinate expression, this newly defined quantity is symmetric in x and y, obeys a kind of distributive law, and is linear with regard to multiplication of either factor by a number. In the terms that mathematicians apply to such cases, the expression $< x, y >$ is a *symmetric bilinear form*. This designation means that, for all x, y, z, and c, the following formulas hold universally:

$< x, y > = < y, x >$
$< x, y + z > = < x, y > + < x, z > < x + z, y> = < x, y > + < z, y > < cx, y > = c< x, y > < x, cy > = c< x, y >$

All of these can be seen from the coordinate expression of $< x, y >$.
These are the defining properties of a *symmetric bilinear form*. If the first property doesn't hold, it's still a *bilinear form*, but not a *symmetric* one.

Finally, one other observation follows from the coordinate expression of $< x, y >$. One has

$$< x, x > = x_1x_1 + x_2x_2 + \ldots + x_nx_n = x_{12} + x_{22} + \ldots + x_{n2} = |x|^2 \geq 0$$

In other words, $< x, x > \geq 0$. Furthermore, $< x, x > = 0$ if and only if $x = 0$. A bilinear form with these two additional properties is called *positive definite*.

Therelationshipof$< x,y>$totheangle between the vectors x and y was established for vectors lying in the Euclidean plane. But any two vectors in R^n determine a plane. In giving distance a universal significance in R^n, one gives a universal significance, as well, to any other measurement that can be defined in terms of distance. But this applies, in particular to the measurement $< x, y >$, commonly referred to, in this context, as an *inner product* or the *inner product of the vectors x and y*. One therefore, thinks, in general, of the expression

$$\cos\theta = < x, y > /|x||y|$$

as determining, or measuring, the cosine of the angle between any two vectors x and y in R^n.

The case $\theta = 90^0$ is particularly important because its cosine, $\cos(90^0) = 0$. If follows from the formula that two vectors are orthogonal, are perpendicular, if and only if their inner product is zero.

As a final cultural point: the modern way of looking at an inner product is that it constitutes a structure that one has *added* to the vector space. Yet, such a perspective has it backwards. One does not create the reality that one measures; one creates means of *measuring* reality. To introduce an inner product on a vector space is, fundamentally, to recognize a measurable feature, and to introduce a way of measuring that feature, that one had previously been treating as an omitted measurement.

To introduce and measure a new distinction is, at bottom, to *recognize* an additional distinction. It is to treat, as relevant, a feature that had previously been treated as irrelevant.

[1] Ayn Rand, *Introduction to Objectivist Epistemology Expanded Second Edition*, In the section "Cognition and Measurement," p 7 in the paperback edition

[2] Archimedes, *The Works of Archimedes*, 1897, Cambridge: at the University Press. See "On Equilibrium of Planes," pp 189-220 and "On Floating Bodies," pp 253-300

[3] Israel Kleiner, A History of Abstract Algebra, 2007, Birkhauser, Boston, p 79

[4] Archimedes, “On Equilibrium of Planes,” first postulate in Book I, p 189, Following the unfortunate Euclidean practice, Archimedes does not offer a discussion of his postulates

[5] Archimedes, “On Equilibrium of Planes,” This interpretation of the first postulate is implicit in the argument for Proposition 6 in Book I

[6] Archimedes, Prop 6-7, p 192

[7] Archimedes, Prop 7, p 193

[8] Archimedes, p 257

[9] Archimedes, p 257

[10] Archimedes, p 263

[11] Harriman, “The Development of Dynamics in Chapter 4 “Newton’s Integration,” p 117-30. Newton’s introduction of the vector concept is discussed on page 118-9. IHarriman explains how Newton’s identification of velocity and acceleration as vector quantities, as magnitudes in a certain direction, were essential to Newton’s discoveries and formulations of his laws of motion

[12] Paul R. Halmos, *Naïve Set Theory*, Springer, New York, 1974, discussion of Cartesian products, in Section 9, “Families,” p 36

[13] Rand, section entitled “The Cognitive Role of Concepts,” p 63

[14] Rand, section entitled “Abstraction from Abstractions,” p 19-28 is relevant here

[15] Rand, section entitled “Abstraction from Abstractions,” p 19-28

[16] Paul R. Halmos, *FiniteDimensional Vector Spaces*, 1958, Princeton, New Jersey, D. Van Nostrand Company, Inc., p 10

[17] Kleiner, p 80-81

[18] Halmos, *Vector Spaces*, p 7

[19] Halmos, *Vector Spaces*, p 9

[20] Halmos, *Vector Spaces*, p 7

[21] Halmos, *Vector Spaces*, p 10 (definition of basis)

[22] Halmos, *Vector Spaces*, proof of Theorem 1, p 13

[23] Halmos, *Vector Spaces*, proof of Theorem 1, p 13

[24] Halmos, *Vector Spaces*, p 10

[25] Halmos, *Vector Spaces*, This point is made on p 14

[26] It is, in part, because of this very uniqueness that mathematicians embrace the set theoretic versions of number that I presented in Chapter 6. Mathematicians, on an abstract level, only care about mathematical structures. If a self-consistent theory can construct the natural numbers and the real numbers that means that their

axioms of the number system are consistent. The rest, they don’t really care about. Mathematicians, at least officially, don’t care whether their mathematical theories have any real-world referents. They care only that they can be derived from a “consistent” set of axioms they have accepted as reasonable.

[27] Kleiner, p 79

[28] Kleiner, p 87

[29] Kleiner, p 80

[30] Kleiner, p 82

[31] Kleiner, p 82

[32] Kleiner, p 85

[33] Halmos, *Vector Spaces*, p 64-68, or, in general, any book on linear algebra or vector spaces

[34] Halmos, *Vector Spaces*, p 55

[35] Halmos, *Vector Spaces*, p 67

[36] Halmos, *Vector Spaces*, p 58-59

[37] Jeremy, Gray, *Plato's Ghost the Modernist Transformation of Mathematics*, 2008, Princeton, Princeton University Press, as noted in Chapter 4, apparently Newton held this view of numbers. On page 134, Gray refers, in passing, to the "Newtonian view that the real numbers were ratios of quantities"

[38] Euclid, *Elements*, edited with notes by Thomas L. Heath (New York: Dover Publications, 1956), Prop I.47

[39] Euclid, Prop. II.12

Chapter 8 Abstract Groups and the Measurement of Symmetry

Symmetry, Similarity, and Measurement

Abstract algebra, algebraic topology, and non-Euclidean geometry helped bury, over a century ago, *the science of quantity*, as a characterization of mathematics.[1] Among the earliest, and among the most important, of such disciplines, and part of abstract algebra, is the theory of *abstract groups*.

There was never a need for this burial; it reflects, in part, an overly narrow view of quantity, one that has not been significantly widened since the collapse of classical civilization. But, to widen the concept, quantity, it is helpful to understand how *quantity* relates to one's conceptual knowledge of the world.

I have argued that *the science of measurement* is a better characterization of mathematics than the *science of quantity*. Mathematics is *about* quantity, but its central concern, the reason we need a science of mathematics, is to establish *relationships* between quantities. Mathematics is a science of method, specifically of measurement.

And *measurement* is all about distinguishing *differences*, of *measuring* differences by relating them quantitatively, specifying differences within an axis of similarity.[2] Mathematics, qua science of measurement, applies to any axis of similarity that can be distinguished and identified by the human mind. It covers, from a separate, differential, perspective, the entire conceptual realm.

Symmetry, the concern of group theory, is a very broad category that relates to the still broader category of similarity.[3]
If one moves a solid, rigid object, such as a pencil or a book, the moved object remains the same object. Its parts retain the same internal relationships. Its shape, in all its aspects, remains the same. What changes is the location of the object and its orientation in space. One says, for this reason, that space is symmetric, with respect to changes in location and orientation. The internal nature of an object, or of a system of objects, does not depend upon its overall

location or orientation.

A reflection is another kind of symmetry. Objects, as such, cannot be reflected, but their images can. And the shape of one object can be a reflection of another, as one's right hand is a reflection of one's left hand.

In thinking about reflection, one thinks either literally or figuratively of the action of a mirror. The internal relationships of a reflected object remain the same, but the reflection is a kind of reversal. A left hand is reflected to look like a right hand. A left hand cannot occupy a space occupied by a right hand; its reflection can. Conversely, a right hand, or an already reflected left hand, is reflected to look like a left hand. Reflection is a toggle: The reflection of a reflection reverses the reversal to the original orientation.

But not all objects look different when reflected. Some objects, objects possessing a certain kind of symmetry, remain the same under reflection. For example, a featureless rectangle remains, upon reflection, a featureless rectangle of the same shape. A perfectly symmetric face looks the same under reflection. In point of fact, every feature of that face participates in the reflection, but the resulting image is indistinguishable from the original. Objects with this kind of symmetry are said to be *bilaterally symmetric*.[4]

Now consider the positions of a cubical die (singular for dice). Distinguish the positions of the die in two respects, namely, which side of the die is up and which side is facing an observer. There are 24 such positions. Why? First, any one of the six sides can face upwards. The opposite side is automatically face-down and any one of the four remaining sides may be facing the observer. For example, if the side with six dots is face-up, then the side with one dot is face-down. The side facing the observer has either two, three, four, or five dots. So that's 4 faces that might be facing the observer for each top face. Four times six, i.e., 24 possibilities.

The dots serve to distinguish the sides of the cube. But the shape and location of the space filled by the cube is the same in all 24 positions. These 24 positions are considered the symmetries of the cube (up to reflection); there are 24 different ways that the cube can occupy any particular space.

In one sense each of these 24 alternatives is the same; they are the same insofar as they all occupy the same space. But they are different in another sense in that the arrangement of the faces is different: the same, but also, within this sameness, different.

Finally, consider a deck of cards. Any one of 52 cards may be on top; any one of the remaining 51 cards may be next, *etc.* Accordingly, there are $52 \times 51 \times \ldots \times 2 \times 1$ (= 52!, read 52 factorial) possible orderings of the cards. (One uses the

notation *52!*, by definition, to designate this product of the integers from 1 to 52.) From the backs of the cards, these orderings are indistinguishable; but from their faces there are 52! different orderings.

Once again, there is a kind of symmetry, a respect in which each alternative is the same. Yet, at the same time, there is a respect in which every possibility is different.

To measure symmetry, or, more precisely, to determine a system of symmetry measurements, is to to identify and to track, the respects in which these possibilities, differences within sameness, differ from each other. But not to measure them in the conventional way, in terms of all the other characteristics of the objects involved. Measurement of symmetry zeros in on a specific set of potential differences within a wider similarity. For example, one is concerned specifically with 24 distinguished positions of a die without regard for, say, the size or material composition of the die, or the *extent* to which a particular face is facing the observer.

A characterization of symmetry, attributed to Hemann Weyl, author of the classic, *The Classical Groups*, is close to the mark. “A thing is symmetrical if there is something you can do to it so that after you have finished doing it it looks the same as before.”[5]

This is certainly an apt characterization. But the essence of symmetry, I believe, is deeper than this. Most generally, as I have illustrated, symmetry involves situations in which various alternatives are identical from one particular perspective, yet different from a second, equally valid, perspective. This characterization of symmetry applies, universally, whenever one isolates a dimension or a specific constellation of dimensions along which similar things differ from one another

The purpose of mathematical groups is to measure symmetry. But a group perspective, as a system of measurements, differs from, yet supplements, the determinations of other systems of measurement.

For example, when a number is applied to measure a magnitude, it specifies that *magnitude* by identifying its relationship to a standard. Alternatively, when a number is treated as a ratio, that number measures the *relationship* between two magnitudes, without regard to a choice of standard or, indeed, to the particular kind of magnitude to which it may apply. To say that A is twice B is not, per se, to identify either of the physical magnitudes, not even in regards to the *type* of magnitude under consideration.

The difference is one of perspective, of the aspect of the situation that is being identified. In either application, a number specifies a relationship between two

magnitudes. In the first case, the numbers are used to specify one *particular*, as opposed to other particulars, by specifying its *relationship to a known magnitude*. In the second case, the numbers are used, specifically to specify a *relationship*, as opposed to other possible relationships between two magnitudes. The numbers are used to measure symmetry, to measure, to distinguish and specify, the *respects in which two similar quantities can differ*. In this respect, as we shall see, positive real numbers function as a mathematical group.

The precise focus of a group-theoretic perspective on symmetry will become clearer as we proceed. The general purpose of this chapter is to show how abstract groups arise, what they mean, and what they measure. I will explain the key concepts and outline the transition from *transformation* groups to *abstract* groups.

The most profound and influential early development in "group theory" was contributed by a young mathematician by the name of Galois, in 1832, the night before his tragic death in a duel. He began with a most prosaic unsolved problem regarding fifth degree polynomials: Can one find a general formula, in terms of radicals (e.g., fifth roots) involving the coefficients of the polynomial, to solve a polynomial equation of degree 5?

Such an equation can be written in the form: $Ax^5 + Bx^4 + Cx^3 + Dx^2 + Ex + F = 0$. One says that this is a polynomial of degree 5 because 5 is the highest exponent of x.

One seeks a general formula for a solution x in terms of the numerical coefficients A, B, C, D, E, and F. The formula for solutions of polynomials of degree two, the celebrated quadratic formula for polynomials of the form $Ax^2 + Bx + C = 0$, is taught to middle school and high school students today. Important special cases of the quadratic formula were known, in somewhat different forms, in antiquity. Since the Renaissance, formulas had been discovered, as well, for polynomials of degrees three and four. But a formula for degree five had proven elusive.

Enter Galois. Facing the threat of the impending duel the next morning, concerned, evidently, with the possible consequences of the duel, he committed certain of his discoveries, his lasting legacy, to the written record. These discoveries proved that a general formula for solving degree-five polynomials does not exist. But Galois' method of discovery is what made him immortal. For it helped spark a new branch of mathematics, one of fundamental importance: group theory. By studying the *symmetries* of the fifth degree polynomial from the perspective of its roots, he was able to cut through the inessentials and get at the essence of the problem without becoming mired in the kind of calculations

that would otherwise have been necessary, that had been common to earlier attempts.[6] Galois' solution was a *conceptual* breakthrough.

Group theory is useful whenever symmetry is important. In general, mathematical groups provide a way to exploit the ways that various aspects of things are the *same* without losing sight of their *differences*. Regarding these differences, one is specifically interested in the *scope* of these differences and in the *structure* of the relationships within the system of measurements that distinguishes them.

For example, as it turns out, a rotation of physical objects is captured mathematically by a certain kind of linear transformation. (Regarding linear transformations, see Chapter 7.) Such a linear transformation identifies a relationship between the two positions of the object. But, from a symmetry perspective, one is interested, not only in what these linear transformations differentiate, but also in what they preserve.

Namely, one is interested in the fact that the *shape* of the transformed object, as manifested in the dimensions of its parts and the angles between its parts, is unaffected by the linear transformation. One isolates this shape-preserving group of linear transformations as, specifically, that subset of linear transformations that preserve the shapes of the objects that they transform. And, insofar as one is interested in *symmetry*, one focuses on the *structure* of this particular system of linear transformations without regard to other more specific characteristics of any specific object that might be rotated according to the specifications of the linear transformations.

Symmetries are abundant in both the natural and the manmade world. And wherever symmetry exists, there is a context and a perspective from which the symmetry matters. For example, because crystalline structures are symmetric, group theory plays an essential role in their study. Because of certain subtle symmetries in the way one measures spatial and temporal relationships, group theory also informs more theoretical pursuits such as quantum mechanics and General Relativity. Within mathematics, itself, symmetry considerations crop up in almost every mathematical specialty. Group theory has been dubbed "the study of symmetry" and group theory is *the* mathematical tool required for its study.[7]

But does group theory study quantity, once considered the subject matter of mathematics? To accommodate group theory, do we need to broaden our understanding of mathematics, beyond quantity, as a discipline or do we, perhaps instead, broaden our understanding of quantity and measurement? Does group theory force a change in the paradigm of mathematical pursuits or does it

simply study a new kind of quantitative relationship unknown or unrecognized in antiquity? Should we broaden our view of mathematics and leave quantity behind? Or do we broaden our view of mathematics by broadening our view of quantity and measurement? And is this a specifically scientific question or is it a philosophical one? To address this last: Historically the answer was made on philosophical grounds. And, perhaps, it is the philosphical underpinnings that need to be challenged. Conversely, if one rejects those philosophical underpinnings, one should reconsider, as well, the modern views of mathematics that they spawned.

A central theme of this book has been that mathematics is the science of measurement. Group theory is part of that science and it measures something in the world; it measures quantity. Just as numbers relate magnitudes and provide the means to relate a particular magnitude to a standard, a system of symmetry measurements relates similar objects and thereby, provides the means to specify a particular point along a symmetry spectrum by relating it to a chosen standard point on that spectrum.

What is a mathematical group? The standard answer is to simply offer a definition and then to show how various examples satisfy the definition: First the definition and then the examples. And such examples, typically, do not function to motivate the concept, but, rather, to establish the existence of groups satifying the stated definition. The better textbooks, indeed, use well-chosen examples to motivate the *development* of the theory, but the primary use of examples is typically to motivate and develop key *theorems*, important truths, regarding the subject. The *concepts* required to express and prove these theorems usually receive less attention and motivation even in these better treatments.

Motivating theorems is important, but motivating concepts should come first. Motivating mathematical concepts, showing how they arise, what they integrate, why they are important, and how they relate to the world, should have central importance in learning, teaching, and thinking about mathematics.

The goal of this chapter is conceptual; it is not to systematically teach group theory. Rather it is to demonstrate the way that mathematical groups relate to the world and to elucidate what they measure. Necessarily this requires mathematical content. But I want, as far as possible, to explain group theoretic concepts for a non-mathematical audience. So my approach will be to start with a simple, if somewhat artificial situation and develop key concepts, in the simplest possible way, from their base in perception.

The Puzzle-Piece Transformation Group

I start with a simple problem, depicted in Figure 1: a puzzle block with one equilateral triangle as its puzzle piece. The front of the piece is green (slant pattern); the back, red (brick pattern). The piece can be put in upside down showing the red back of the piece. How many ways will the piece fit the puzzle?

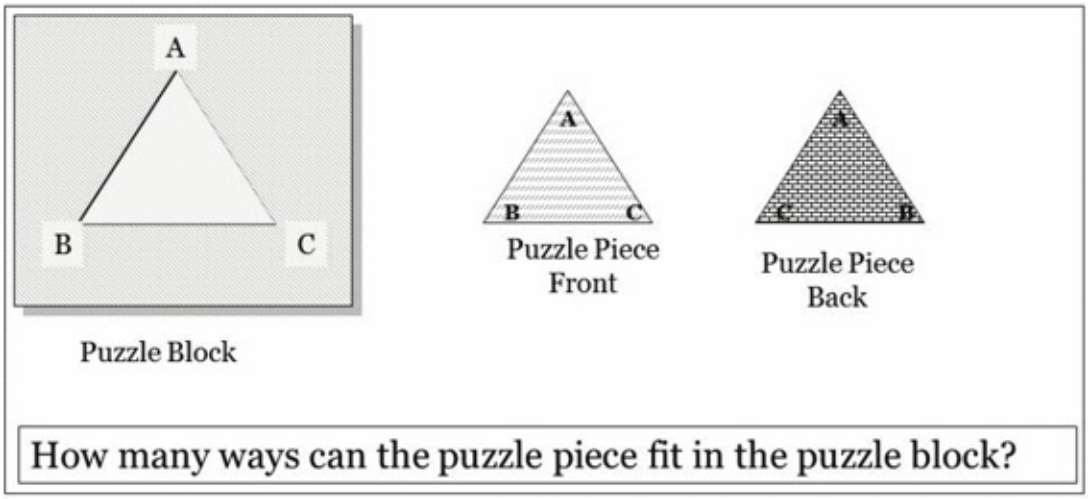


Figure 1

As Figure 2 depicts, the answer is six. The corner of the triangle labeled A can be matched with any of the puzzle-block vertices and then B can be matched with either of the remaining vertices, leaving C to match the remaining vertex. The position of A, then, has three choices. For each of these choices, B has two possibilities. So the number of possibilities is $3 \times 2 = 6$. I will sometimes follow standard practice and refer to these six positions as the *symmetries* of the equilateral triangle.

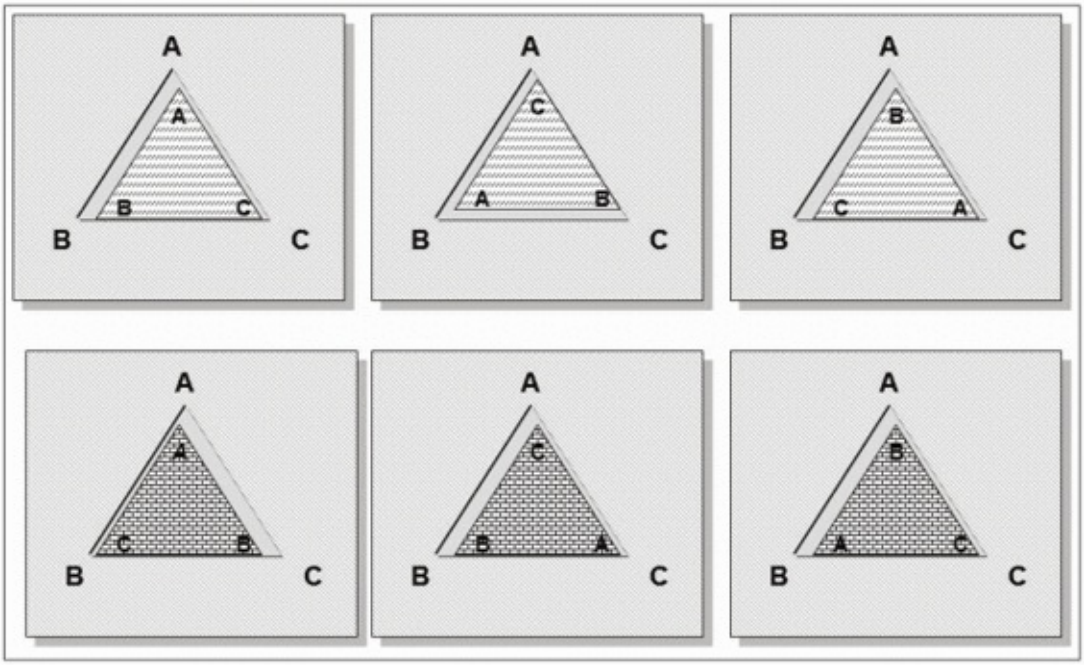


Figure 2

Notice something about this. First, the multitude of choices depends entirely on the symmetry of the triangle, on its three equal sides and also the fact that its shape is the same viewed from the back as from the front. The triangular piece has the *same* shape from each of these six perspectives.

Yet, at the same time, there *are* six different perspectives. Each perspective represents a distinct positioning of the triangular piece. The positions are different, but they can also be regarded as the same or as similar insofar as they occupy the same shape in the puzzle block. As Ayn Rand analyzes similarity in regards to concept formation, the difference between two similar things is one of measurement.[8] In this case, the similar things consist of six different *orientations* of the triangle within the same space. The relationship between any two of these orientations is a quantitative relationship and a specification of this relationship with respect to a standard initial position is a measurement.

What are those quantitative relationships?
If the position in the upper left of Figure 3 is taken as the starting position, the other two triangle positions along the top are related by a rotation from the starting position (labeled E). Similarly, the bottom three positions each can be reached by reflecting the triangle from the starting position along one of its axes. I have given each of these relationships a name in Figure 3. In the top row, “R”

stands for a counter-clockwise rotation that sends each vertex to the next available vertex in the puzzle block. The second rotation next to it is called, suggestively, "R^2". The bottom three positions all relate to the starting position by a reflection. As depicted, these reflections labeled A_r, B_r, and C_r differ, respectively, from the original position, by a reflection on the axis through vertex A, B, or C. Thus A_r is the reflection about the vertical axis that passes through the vertex A.

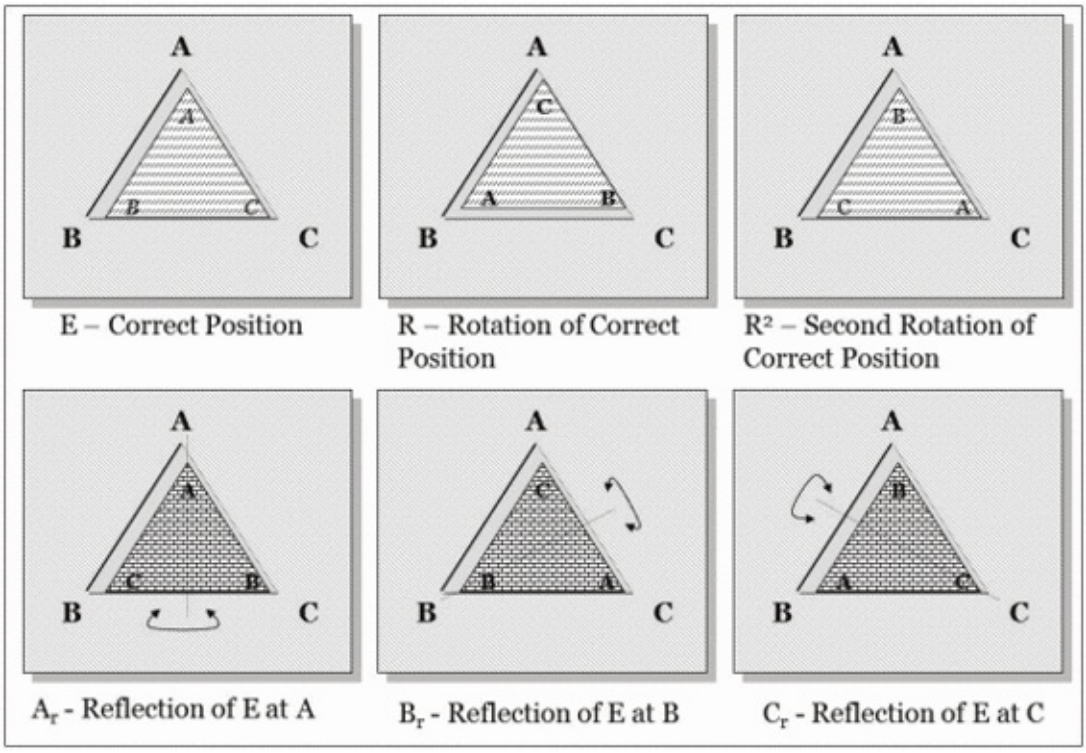


Figure 3

The names for the three reflections may appear intimidating, but I have named them that way for a reason. First, the capital letter is there to remind us which vertex, relative to the puzzle block is being kept fixed. Thus, A_r reminds us that the vertex at the top, in position A, is to be kept fixed. The subscript r after the capital letter stands for "reflection" and is there to remind us that the relationship is a reflection.

Notice that in identifying these relationships there are a great number of measurements, specific to any individual case, that I am ignoring. For example, I ignore the size of the puzzle and the puzzle piece. I ignore the speed with which one might move the piece from one position to another. I even ignore the fact that R is a rotation of 120^0. I focus solely on the specification of the change from

one position to another and only in regard to the change in the positions of each vertex. As in all such cases, to ignore other measurements is not to pretend that these relationships do not exist; it is, rather, to recognize that these other relationships do not affect the relationship under investigation.

Most importantly, one can look at each of these transformations (rotation or reflection) as acting in a way that doesn't depend on a particular initial starting position. No matter what position a triangle occupies in the puzzle block, it will be moved to a new position by following one of the recipes or prescriptions identified in indicated in the diagram (Figure 3). No matter what triangle vertex occupies puzzle-block corner A, the transformation R will rotate it to corner B. Similarly, A_r will interchange the triangle vertices located at puzzle-block positions B and C, leaving alone the triangle vertex located at puzzle-block position A. The transformation labeled E can be regarded as the trivial transformation that leaves everything the way it already is.

Finally, these transformations can be applied in sequence. For example, Figure 4 shows the effect of following the rotation R by a reflection about the axis through A. Since Figure 2 contains all possible symmetries of the triangle, the combined effect of R followed by A_r must be one of those symmetries. So the combined net effect of the two transformations must be identical to one of the transformations depicted in Figure 3. And, indeed, the rightmost triangle in Figure 4 can be derived from the leftmost triangle by a reflection about the axis through B. So the combined effect (see Figure 3) is B_r.

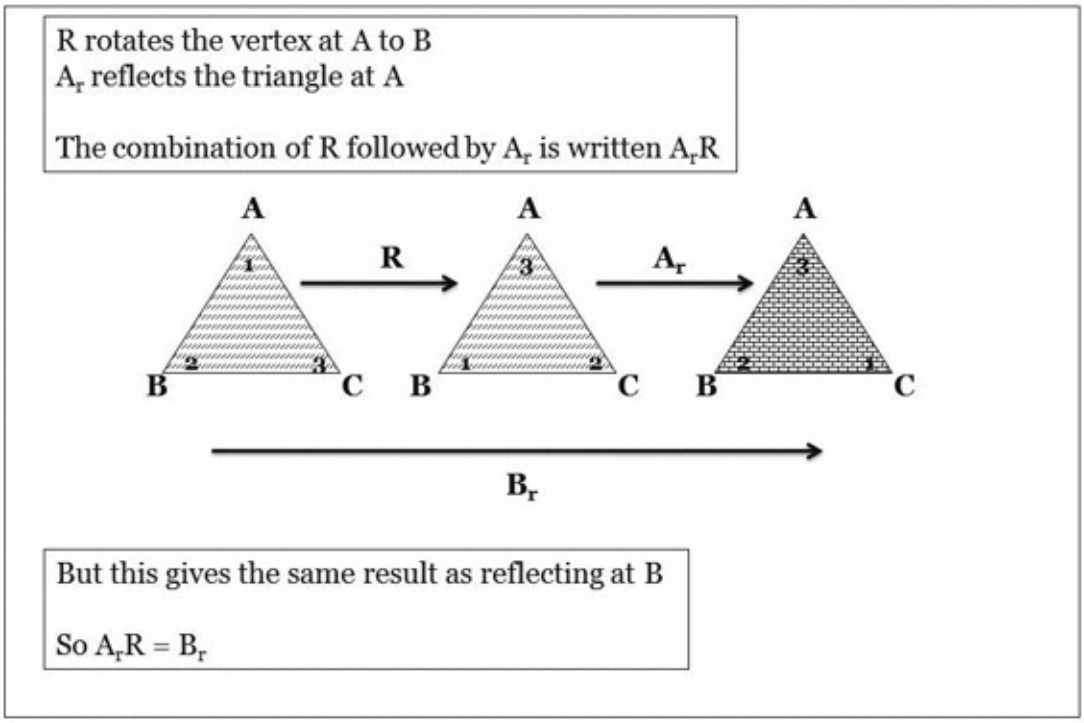


Figure 4

The combination of R followed by A_r is normally written A_rR. Notice that the letters are written in reversed order from what one might expect. Mathematicians follow this practice because each transformation is thought of as acting on something, something that, if shown explicitly, would be represented to the right of the identifier of the transformation. So the rightmost transformation in the expression acts first, followed by the transformation to its left.

This process of combining two transformations to yield a third transformation is analogous to multiplication. Indeed, it is usually called “group multiplication” and is also completely analogous to the multiplication of matrices and linear transformations discussed in Chapter 7. The relationship that I have just identified is written $A_rR = B_r$. Also, anticipating definitions to follow, I will refer to particular transformations as *elements of the transformation group*, or as *group elements*.

This multiplication relationship is further dramatized in Figure 5. In the top row, the transformation on the leftmost triangle is thought of as occurring in two steps. In the bottom row, one is interested only in where the triangle ends up. The precise way that the triangle got there is unimportant. B_r identifies the

composite transformation that transforms the triangle to its end state.

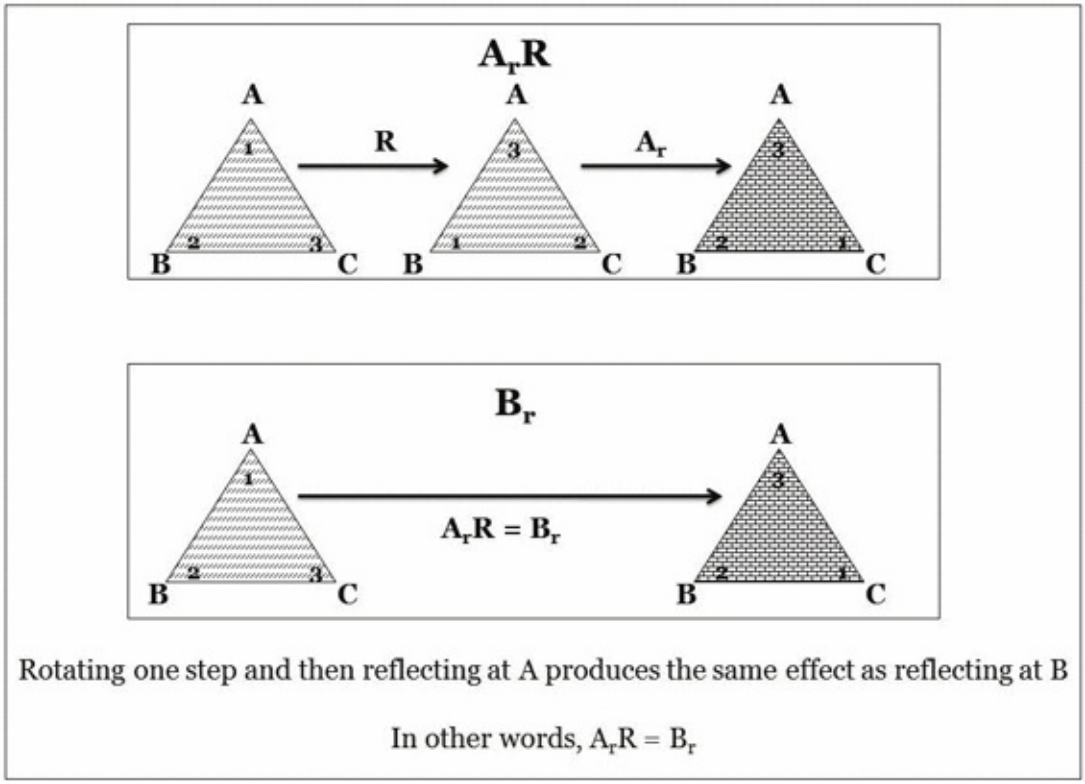


Figure 5

In sum the elements of the transformation group express the *quantitative relationships* between any two positions of the triangle within the puzzle block. From a transformation group perspective, one's measurement of these relationships is no more and no less than a specification of the starting and ending positions of each vertex, as a complete specification of the differences of interest. All other measurements that might characterize a concrete instance are omitted as unimportant from a symmetry perspective.

Taken together these group elements possess their own arithmetic because they can be thought of as operating to transform one position of the puzzle piece to another. When interpreted in this way these elements serve as a closed system of measurements that can, in the sense I've indicated, be multiplied together to yield other measurements within the same system.

The analogy of group multiplication to multiplication of numbers can be illuminated as follows: Consider a "scaling transformation" of positive numbers that consists of multiplying every number by a fixed positive scaling factor. For

example, if the scaling factor were 3, the scaling transformation would multiply every number by 3. With a scaling factor of 3, the number 5 scales to the number 15, the number 20 scales to the number 60, and so on.

Every positive number can act as a scaling factor. So, for example, a scaling factor of 2 scales the number 5 to the number 10 and the number 20 to the number 40. To apply a scaling factor to a number is to multiply that number by the scaling factor.

A scaling factor can be viewed as a transformation of the set of numbers. Moreover, in applying a scaling factor, one transforms each number to a different number while *preserving the ratio* between every pair of numbers in the set. Suppose, for example, that one applies the scaling factor is 2, to the numbers 20 and 5. The ratio between them is $20/5 = 4$. The scaling factor maps them, respectively, to 40 and 10 and the ratio is, again, $40/10 = (2 \times 20)/(2 \times 5) = 20/5 = 4$. It is in this sense, in the sense that the scaling factors preserve ratios, that the group of scaling factors measures symmetry, a symmetry of the real numbers.

Again we see the general pattern, the aspect in which the transformed object is changed and the aspect in which it remains the same. A scaling transformation changes each number into a different number. But the *set of numbers* is unchanged by the transformation and *the ratio between any two numbers remains the same*.

If one follows the application of one scaling factor by the application of a second scaling factor, the result is a scaling transformation by a third factor. Thus, if one applies, first, a scaling factor of 3 and then, second, a scaling factor of 2, the result is a scaling factor of 6 ($= 2 \times 3$). To multiply 5 by first 3 and then by 2 is the same as to multiply 5 by the product of 3 and 2.

In this instance, multiplication of scaling transformations is the same thing as multiplication in the usual sense. The example of scaling transformations, then, captures precisely the relationship of the expanded concept of “group multiplication” to the usual arithmetic concept of multiplication. Group multiplication is an extension, a generalization of the usual concept of multiplication that captures one important aspect or facet of ordinary multiplication, namely, multiplication considered as a transformation. In, exactly, this respect, as I also discussed in Chapter 7, a composition of linear transformations is a natural generalization of the arithmetic concept of

multiplication.

Transformation Groups

Let us now return to the puzzle piece transformations. When a group (the next level of abstraction, yet to be discussed) arises as a system of transformations, it is called a *transformation group*. In studying a transformation group, one can selectively attend either to the action of its elements on the *object* it transforms or upon the *relationships* between the transformations. Specifically, the multiplication of transformations that I have just exemplified possesses certain properties that are universal to all transformation groups. These properties, as I will elucidate, are all embodied in the puzzle-piece example.

First, as already elucidated in the extended example, one can multiply any two transformations to get another transformation. And notice that, in general, any property of the transformed object that is preserved under any two transformations will also be preserved by the product of those transformations. If a transformation A preserves property Z and transformation B also preserves property Z then A followed by B also preserves property Z.

Next, consider what happens in a series of three transformations, as depicted in Figure 6. In Figure 6, R is followed by A_r, which is followed by C_r, resulting in a composite transformation that one can identify as R^2 (as shown in Figure 3):

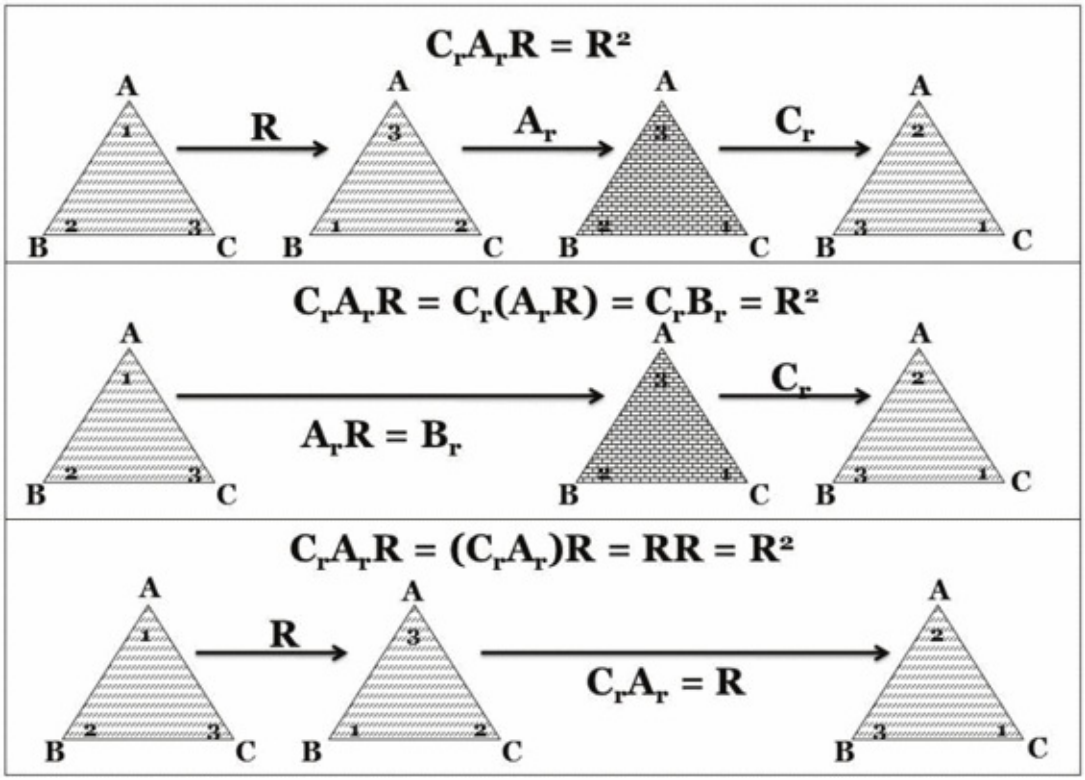


Figure 6

One can *analyze* this sequence of transformations in three different ways. To begin, in the first row, one performs all three transformations in succession without making any assessments along the way. One can see that the resulting transformation is R^2 simply by comparing the last triangle on the right with the first triangle on the left.

The second row depicts the same series of transformations. However, this time one stops to observe the effect of the first two transformations, before applying the third. One observes that the product A_rR of the first two transformations is equal to B_r. In effect, to multiply the three group elements, one first calculates A_rR and then multiplies the result by C_r. One represents this order of calculation by $C_r(A_rR)$. So one has $C_r(A_rR) = C_rB_r = R^2$. Although the *calculation* is different, the *series of transformations it captures is the same* so the result is the same, as well.

Finally, in the third row, one takes the product of the last two transformations, which turns out to be R. Once this assessment has been made, one begins with the transformation on the left and follows it by the product of the other two. As

before, this is simply a third way of looking at the sequence in the top row. One has $(C_rA_r)R = RR = R^2$.

Putting this together, one observes the *associative law*: $C_r(A_rR) = (C_rA_r)R$.[9]
Now this fact, the associative law for transformations, is completely general, applying to any set of transformations of anything whatever. Nothing in my argument was specific to the example at hand. For any transformation group with group elements A, B, and C, the equation (AB)C = A(BC) holds universally and is called the associative law. The associative law is taken to be one of the defining characteristics of a mathematical group.[10]
The same law applies, under the same name and, for much the same reason, to ordinary addition and multiplication. For example (2×3)×4 = 6×4 = 24. But one could also multiply 3 and 4 first: 2× (3×4) = 2×12 = 24. When multiplying a series of numbers, it doesn't matter which multiplication is performed first. The same is true for the elements of a mathematical group.
There are two other defining properties that define a group. The first is that there exists an element E, called the "identity," such that, for any other group element A, AE = EA = A. The identity plays the same role that 1 does when one multiplies numbers and that the identity matrix does in matrix multiplication. We have already met such an element in the present context: the transformation labeled E that leaves the triangle unchanged. If one either follows or precedes a transformation A by E, the application of E will have no effect on the outcome, will not affect the final state of the triangle no matter when it is invoked: Because the transformation E never changes anything!
One remaining property: The existence of an *inverse*. If A is a group element, A^{-1} is called the inverse of A and is defined as the unique group element such that $AA^{-1} = A^{-1}A = E$. (If A were a number, this inverse would usually be written 1/A. So, for example, the inverse of 5 is 1/5.) The inverse of the rotation R is illustrated in Figure 7.[11] In general, A^{-1} is the transformation that undos or reverses the effect of A.
This, again, recalls matrix algebra. The inverse A^{-1} of an invertible matrix A is characterized by $AA^{-1} = A^{-1}A = I$ where I is the identity matrix.

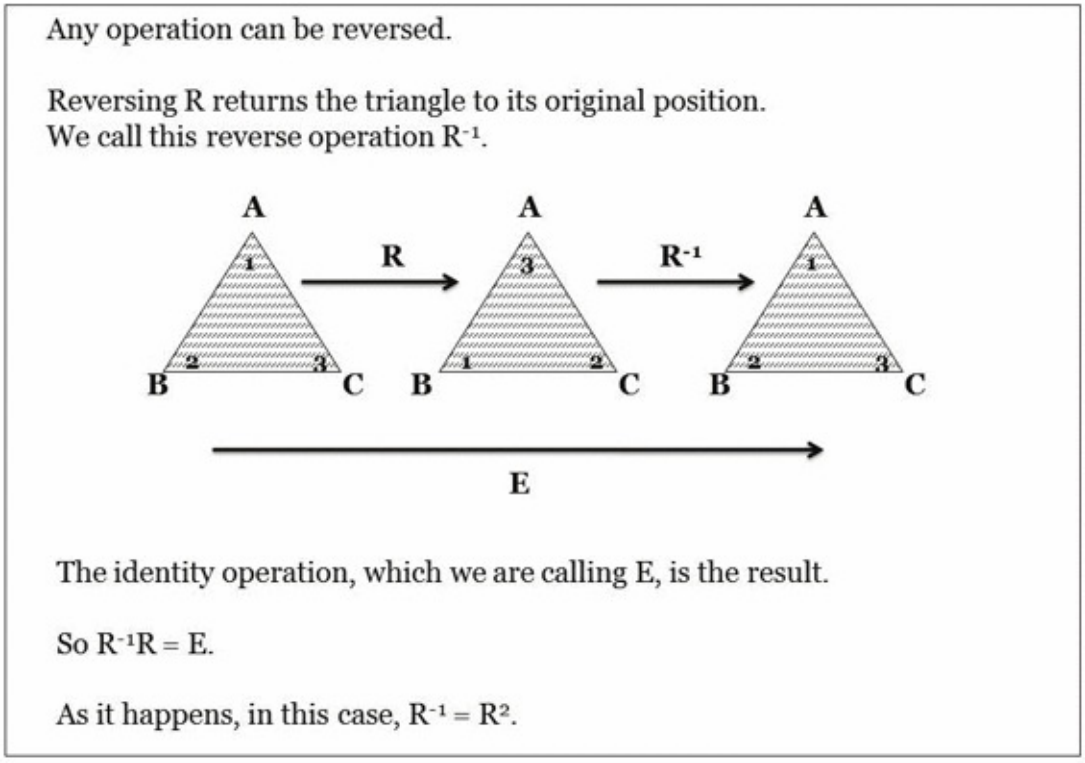


Figure 7

In sum, a set of transformations forms a transformation group G if:[12]

1. The product of two transformations A and B contained in the group G is also contained in G. One designates this "multiplication" by AB. It is the transformation that first applies B and then applies A. This is known as the *closure* principle. One says, also, that the group is *closed* under multiplication.

2. There is a transformation E contained in the group G such that, for any other transformation A contained in G, EA = A and AE = A. This is the transformation that leaves the transformed object unchanged.

3. For any transformations A, B, and C in the group G, (AB)C = A(BC): If one computes the product by multiplying A and B and then multiplying the result by C, the result is the same as first multiplying B and C and multiplying the result by A.

4. For any transformation A in the group G there is another transformation A^{-1} in G that acts in reverse to undo A. So $AA^{-1} = A^{-1}A = E$.

One final word of warning: In general the result of a multiplication of two transformations depends upon the order in which they are applied. For example,

in Figure 8, $A_rR = B_r$, but $RA_r = C_r$. In other words, multiplication for this group is not “commutative”. Or, to put it another way, its elements do not “commute” under multiplication.

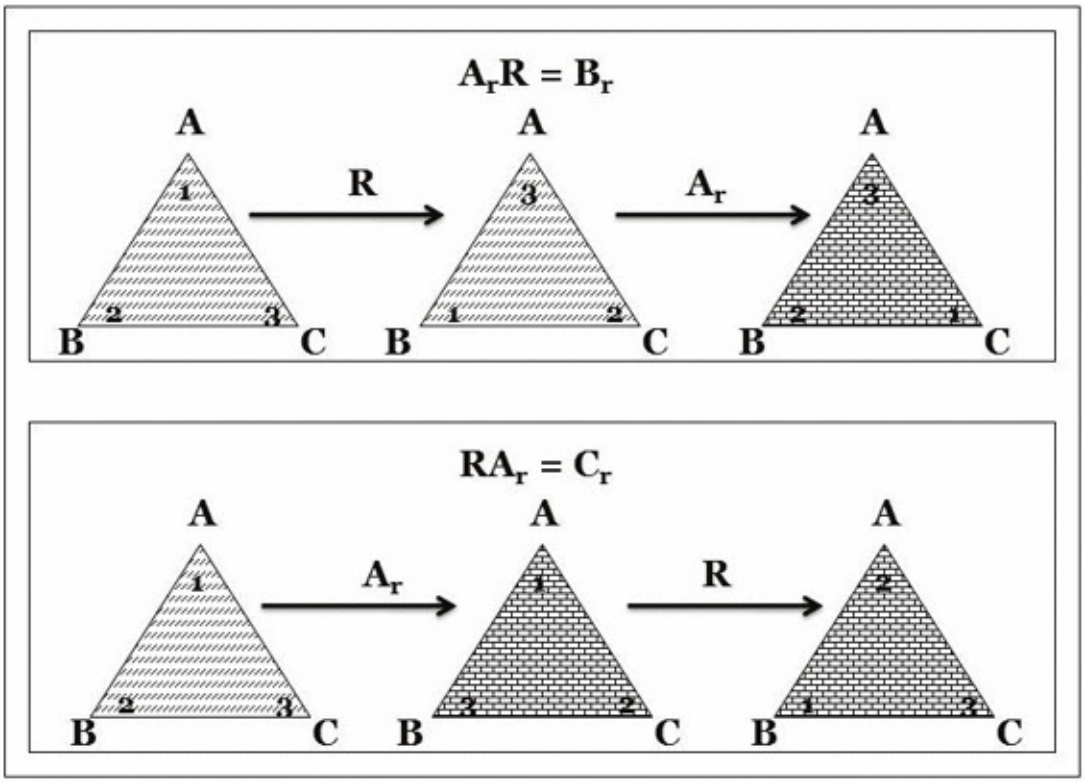


Figure 8

For reference, the group multiplication for this transformation group is captured in Figure 9 as a *times table*:

×	E	R	R^2	A_r	B_r	C_r
E	E	R	R^2	A_r	B_r	C_r
R	R	R^2	E	C_r	A_r	B_r
R^2	R^2	E	R	B_r	C_r	A_r
A_r	A_r	B_r	C_r	E	R	R^2
B_r	B_r	C_r	A_r	R^2	E	R
C_r	C_r	A_r	B_r	R	R^2	E

Figure 9

To read this times table: If an element in the shaded column multiplies, from the left, an element in the shaded row, the result of the multiplication can be found in the intersection of the row and column of the factors. Thus, $RC_r = B_r$. As illustrated in Figure 10, this can be read from the table by following the row labeled R to the column labeled C_r. The answer, B_r, is provided in the cell in which the row and column intersect.

×	E	R	R^2	A_r	B_r	**C_r**
E	E	R	R^2	A_r	B_r	C_r
R	R	R^2	E	C_r	A_r	**B_r**
R^2	R^2	E	R	B_r	C_r	A_r
A_r	A_r	B_r	C_r	E	R	R^2
B_r	B_r	C_r	A_r	R^2	E	R
C_r	C_r	A_r	B_r	R	R^2	E

Figure 10

Subgroups and Quotient Groups

Now suppose one was not able to turn the puzzle pieces over. Suppose, for example that the triangle had a handle on the top that prevented this. Then the reflection transformations would no longer be possible. Every valid transformation would leave the green side showing.

One can see in Figure 3 that the transformations that keep the green side showing are E, R, and R^2. It is easy to check that the product of any two of these transformations will be another transformation that keeps the green side showing. But it is also logical. If A keeps the green side showing and B also keeps it showing, then when I apply them in turn, as BA (first A then B), then each step along the way keeps the green side showing so the green side is still

showing when the series is complete.

This smaller set of transformations is called a “subgroup” in relation to the larger group of transformations. It also has its own arithmetic: EE = E, ER = RE = R, $R^2E = ER^2 = R^2$, RR = R^2, $RR^2 = R^2R = E$ and $R^2R^2 = R$. One can array these in a “times table” as shown in Figure 11. The product of these transformations in unaffected by whether one regards them as belonging to the subgroup or to the full puzzle-piece transformation group. Accordingly, this times table is a subset of the table for the full puzzle-piece transformation group. The smaller table consists of those combinations of rows and columns of the larger table that pertain to multiplications among E, R, and R^2.

×	E	R	R^2
E	E	R	R^2
R	R	R^2	E
R^2	R^2	E	R

Figure 11

Again, to read this times table: If an element in the shaded column multiplies, from the left, an element in the shaded row, the result of the multiplication can be found in the intersection of the row and column of the factors.

Notice that, for this particular subgroup, multiplication is commutative.
Yet another group can be associated with the transformation group of the puzzle piece. Let us return to the original premise that one can turn over the puzzle piece. But now suppose that one *only cares* about which side is showing.
Of the six transformations of the puzzle-piece transformation group, three of them (E, R, and R^2) will leave whatever side is showing unchanged, but will possibly rotate the puzzle piece. The other three are the three reflections that will reverse the side that is showing. Consider, first, the rotations. From the perspective of only noticing or caring which side is showing, all rotations are *equivalent* because none of them affect which side is showing. All of them

preserve the only attribute one now *cares* about: which side is visible. Let us use the letter *e* to denote a transformation that preserves the side that happens to be showing. Because e does not change anything that one cares about, it will function as the identity transformation.

From *this* perspective, E, R, and R^2 are the *same* transformation, namely the transformation that I have just now designated by the letter e. The rotation that might be involved in any of these transformations is regarded as an omitted measurement.

To understand this better, notice that the situation for the full puzzle-piece transformation group exactly parallels this case. Thus, for the puzzle-piece transformation group, one ignores such issues as the size of the puzzle piece or how fast one changes that piece from one position to another, as irrelevant to one's specialized concern: One studies aspects of the situation that do not depend on these details.

By the same token, in this new perspective, the rotational state is now a detail. One may as well forget about fitting the triangle into the puzzle piece at all, since this is no longer an issue. Indeed the reflection that turns over the puzzle piece would apply, equally well, to *any* jigsaw puzzle piece, regardless of its shape.

One is now studying aspects of the situation that do not depend on rotational state. In the original puzzle-piece example, the transformations, say, of two triangles of different sizes, fitting into puzzle frames of correspondingly different sizes, and carried out at different speeds, but performing the same rotation or reflection, are the *same t*ransformation. In the current sidereversing example, two transformations by different rotations, but both preserving the side showing are the *same* transformation. To again paraphrase Ayn Rand's formulation regarding concept formation: the reversed triangle must be in *some* rotational position, but it may be in *any* rotational position. One's identification of the position of the puzzle piece, of which side is showing, doesn't depend on the rotational position.

Similarly, still referring to Figure 3, the reflections (A_r, B_r, and C_r) all reverse whatever side happens to be showing. In this respect these three transformations are also equivalent. Each changes the one thing one cares about and they all change it in the same way. If the side is green, it changes to red; if red, it changes to green. Let us use the letter r to denote a transformation that changes the visible side. Once again, the reversed triangle must be in some rotational position, but it may be in any rotational position.

Notice that if I follow a transformation that preserves the visible side with

another one that preserves the visible side the result, no matter which such transformation I choose, will also preserve the visible side. The product of a transformation belonging to e times another one belonging to e belongs to e. One can express this fact by the equation ee = e. Here, the juxtaposition ee is interpreted as group multiplication resulting from following a transformation belonging to e by another transformation belonging to e.

Thus, for example, the transformation R preserves the side showing. So the transformation that we knew as R, is now viewed as the transformation e. Now $RR = R^2$ and R^2 also preserves the side showing. So R^2 is also the transformation e. Thus $RR = R^2$ is simply *an instance* of the equation ee = e, in much the same way that *2 feet plus 3 feet = 5 feet* is an instance of *2 + 3 = 5*. On one level of abstraction, R is a rotation. On the next higher level of abstraction, R is the transformation e that preserves the visible side of the triangle puzzle piece and, therefore, functions as the identity transformation.

Next, if I follow a transformation that preserves the visible side with one that does not or if, conversely, I follow a transformation that does not preserve the visible side with one that does. I have turned the triangle over exactly once. So the effect is to reverse the visible side. Again, it makes no difference which transformation I choose from their respective classes. So, in symbols, one writes: er = re = r.

Finally, if I follow a transformation that reverses the visible side with another reversal, I have turned the triangle over twice, returning it to the original side. So it has the same effect as one of the rotations that preserve the visible side. In symbols, rr = e.

The multiplication rules that I have been itemizing are captured in the times table depicted in Figure 12:

×	e	r
e	e	r
r	r	e

Figure 12

The group that I have just described is itself a transformation group of the

triangle, but it is not a subgroup of the original group. It no longer distinguishes the puzzle piece according to rotational state. Those states still exist, but they are being ignored; all three rotational states are regarded as the same state. One attends to the sidereversing aspect of the transformation but ignores the rotational aspect captured by the rotational subgroup I discussed earlier.

If the original transformation group is given the name G and the subgroup consisting of the rotations E, R and R^2 is given the name N, then the group that I have just identified, consisting of e and r, is called the *quotient group of G by N* and is commonly written *G/N*.[13] One should notice the analogy of this concept of a quotient group to the concept of a quotient vector space introduced in Chapter 7.

The rationale for this designation of *quotient group* involves the following observations:

1. The non-reversing elements of G are the elements of the subgroup N. They are regarded as acting trivially on the triangle because they do not reverse it. These elements are E, R and R^2. They are regarded as equivalent (also called an "equivalence class") because they all have, to repeat, the same effect on the triangle: they do not reverse it. Call this equivalence class 'e'. To repeat, e is the identity element because it acts trivially on the triangle.

2. Choose one of the sidereversing elements, say A_r. Then all of the reversing elements can be written as the product of A_r and a rotation. Thus, $EA_r = A_r$, $RA_r = C_r$, and $R^2A_r = B_r$. Differing, from each other, precisely by a rotation, which itself does not affect which side is visible, these reversals are also regarded as equivalent because they too have the same effect: they all reverse the triangle. Call this equivalence class r.

3. Each equivalence *class* becomes *one* group *element* in the new group G/N. To have the desired effect (preserving or reversing the visible face), one must choose *some* member from the equivalence class, but one may choose *any*.

4. The equivalence class, i.e., the sidereversing effect, of a product of two elements of G depends *only* on the respective equivalence classes of the two elements. This is why I have been able to construct a times table for the new group.

5. To get the quotient group, one divides the group G into classes. Each element in any particular class differs from the others in that class by an element in the subgroup N. That is, elements g and h are in the same equivalence class if and only if there is an element n in N such that g = hn. In this sense, the subgroup N is the basis for the *division*. And so also, for this very reason, the *number of elements* in the quotient group G/N is the number of elements in G divided by the number of elements in N.

6. In effect, the action of the subgroup N is regarded as unimportant, so if two elements g and h of G differ by an element of N (i.e. g = hn for some element n in N) then they have the same effect on the object that they transform. So h, in this sense, is equivalent to g. The transformations (rotations, in the puzzle piece example) by the elements n of the subgroup N are unimportant.

7. The effect of treating equivalence classes as quotient-group elements is to remove the factor that no longer matters (e.g., the effect of a rotation) from consideration, leaving the factor that does matter (e.g., the visible side of the triangular puzzle piece).

When one first studies arithmetic and learns division, one divides a collection of something into a number of smaller collections, each containing the same number of units (the divisor). Then one counts these smaller collections to get the quotient. From the standpoint of the division, one distinguishes and counts the *collections*, but is not specifically interested in the individual members of each collection

Similarly, and in general, when one creates a quotient group from a group G, by a subgroup N, one divides the group G into “equivalence classes”, such that a) each class has the same number of elements as the subgroup N and b) any two elements (g and h) in the same equivalence class differ by an element of N (there exists an n in N with g = hn). Then one regards the classes as constituting the elements of the quotient group, with the subgroup N serving as the equivalence class that constitutes the identity element.

So the group G/N, in a very literal but generalized sense, is the quotient of the group G by the subgroup N. It is what remains when the factor represented by N has been removed.

But one word of warning: The ability to define the quotient group G/N depended

critically on point 4: that the equivalence class of a product of two elements of G depends solely on their respective equivalence classes. In my example this was guaranteed because the action of each equivalence class could be specified in terms of its effect on a transformed object, an effect that depended specifically on, but only on, the equivalence class of each element of G. But this is not true for all subgroups of all groups. One can always form a distinct set of equivalence classes but in general one cannot define a unique multiplication of equivalence classes to form a group G/N. The required condition can be found in any textbook on group theory: The necessary and sufficient condition is that the subgroup N of a group G be a *normal* subgroup, which means that, for any element g of G, and any element n of N, the element $g^{-1}ng$ (group product of g inverse times n times g) is an element of N. Normality is a condition that must be checked in each case.

Mathematicians study quotient groups and subgroups, in part, because subgroups and quotient groups shed light on the structure of the group from which they are derived. Such light is badly needed; the range of possible structures of groups is vast and constructing times tables is only a small beginning in dealing with this complexity. Part of understanding any particular group is to analyze its structure. And a key part of that structural analysis consists in breaking a group into smaller groups: finding and further analyzing its subgroups and its quotient groups.

The Permutation Group

Now consider an apparently unrelated problem. Suppose that one has three marbles that are distinguished by their color. How many ways can they be arranged in a row from left to right? Such rearrangements are known as *permutations*, a concept essential to the mathematical theory of probability, but which generally arises in a variety of mathematical contexts. Suppose that these marbles are colored Apple red (A), Blue (B), and Clover green (C).

Notice the aspect of symmetry. The three marbles are identical insofar as they are all the same kind of object and have the same shape. But they have different colors. When one rearranges them, the collection is still the same and the space that they occupy is still the same. In that respect, the two rearrangements are the same. But the distribution of marbles among particular positions is different. In this example, that difference is made evident by the differences in the colors of

the three marbles.

There is an essential similarity between the permutation problem and the puzzle piece problem. Both involve symmetry, but there are differences in the kind of symmetry involved. In the puzzle piece problem, different rearrangements of the puzzle piece reflect the symmetries of the puzzle piece. But, in the permutation case, different permutations in the three marbles reflect a different sort of symmetry, a symmetry that is inherent in one's ability to regard a collection of objects as being ordered according to some rule, principle, or characteristic. Under a permutation, the collection remains the same collection, and it occupies the same space, no matter how it is ordered. Yet each arrangement is different in that one can distinguish the different orderings of the collection. In this case the ordering principle consists in lining up the marbles from left to right.

The six possible permutations are depicted in Figure 13. As presented, the two permutations below the top left permutation can each be generated from the one above it by moving the last (rightmost) marble to the beginning, in effect, rotating the letters. Thus ABC CAB is the rearrangement going from the permutation on the top left to the permutation below it. There is a pattern to the three permutations on the right, as well: Each permutation on the right can be derived from the permutation on its left by permuting the two rightmost marbles of the permutation on the left.

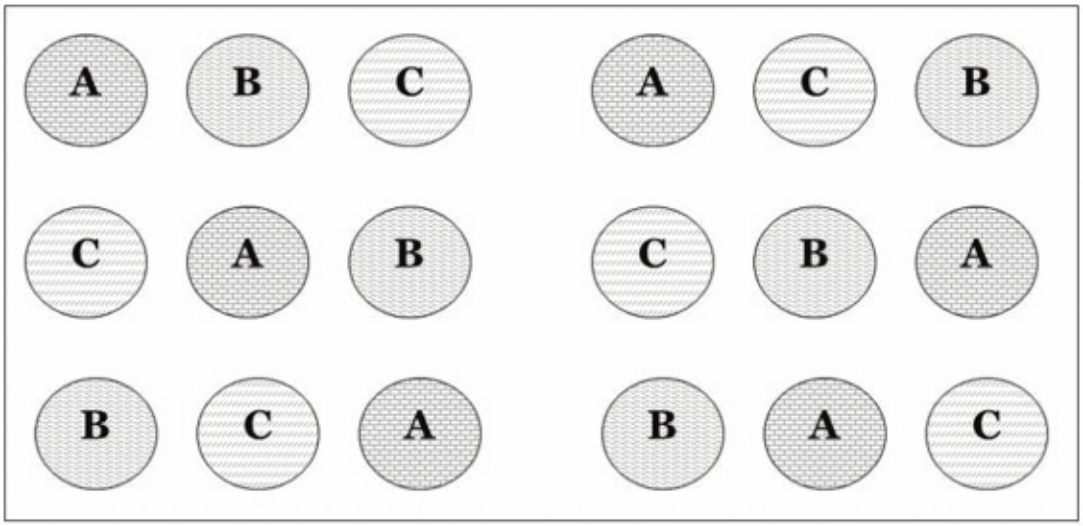


Figure 13

One could now repeat the entire discussion just completed for the puzzle-piece transformation group, but apply it to the "permutation group" that transforms one permutation into another one.[14] But it will save time to simply match the

permutations for the permutation group to the puzzle-piece positions for the puzzle piece group, as in Figure 14.

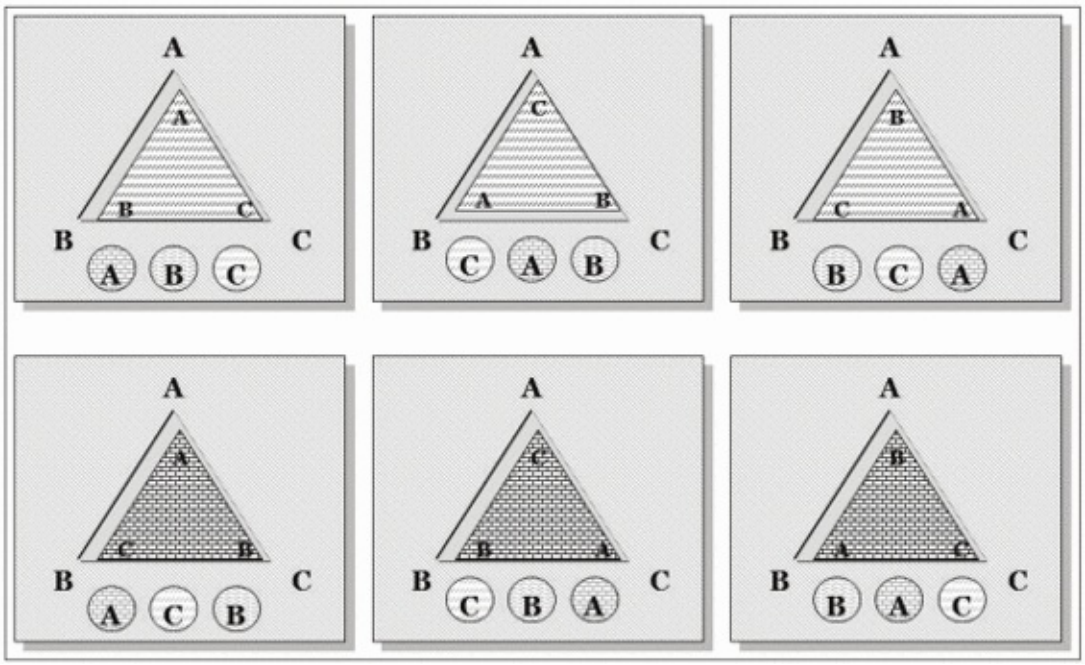


Figure 14

This is an exact match that follows a very simple rule. In each case, the letter at the top of the triangle matches the first marble, the one on the bottom left matches the second marble and the one on the bottom right matches the third marble.

The transformations match, as well, as shown in Figure 15. Rotating the triangle corresponds to moving the third (right) marble in a permutation to the front (left). And a reflection like B_r corresponds to the permutation that leaves the middle marble (B) fixed and interchanges the other two. The descriptions provided under each case can be considered as applying to either the triangle, as compared to the top left triangle, or to the permutation of the marbles, as compared to the top left permutation.

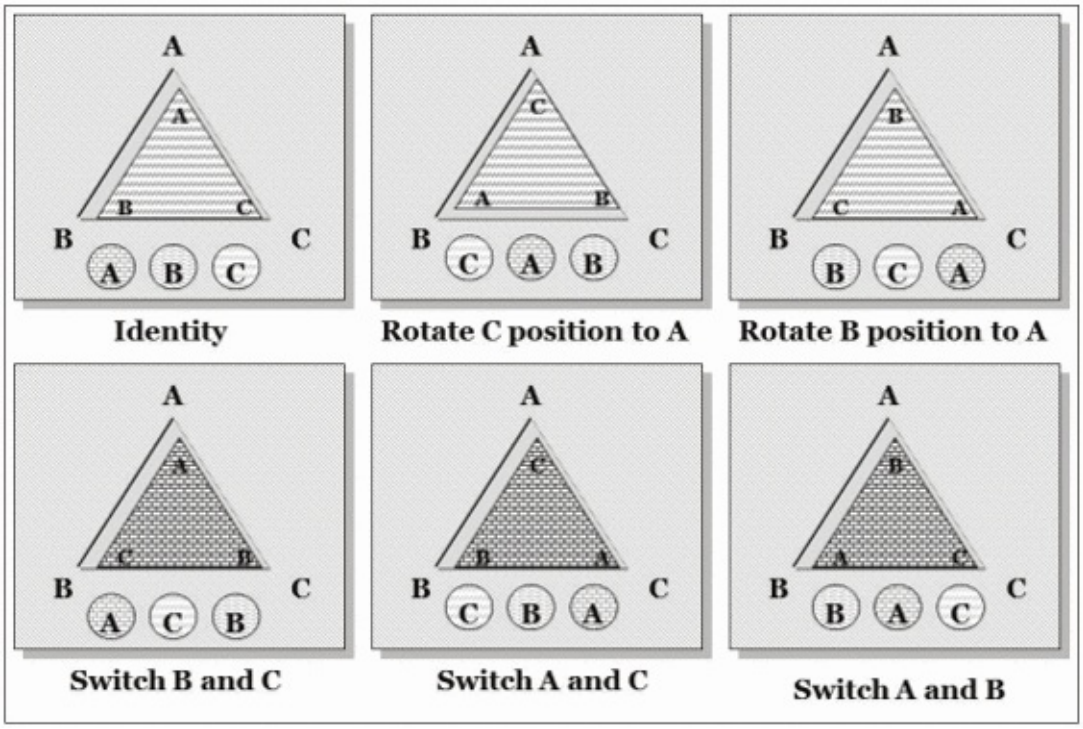


Figure 15

In essence, the puzzle-piece transformation group acts in exactly the same way as the permutation group. If the same names were given to corresponding transformations within the two systems of transformations, the permutation group would have exactly the same times table as the puzzle piece transformation group. The physical objects affected are very different, but the *transformations* of these objects line up exactly. Considered from the aspect of the *relationships between transformations*, of the times tables for the two groups, the puzzle-piece transformation group is identical to the permutation group.

If one omits the particular object of the transformations, on the basis that the group must transform *something*, but may transform *anything* possessing the appropriate symmetries, then one has formed the concept of an abstract group.

A group has an identity element, every element has an inverse, and multiplication is associative. But what is the genus of *group?* What is the sort of thing that possesses a group multiplication with these particular properties?

Mathematicians generally define an abstract group as a “set” possessing a multiplication table that satisfies the properties that I itemized earlier for

transformation groups.[15] As I elaborated in Chapter 6, the word *set*, as used by mathematicians and especially when it is used as a genus, has a highly technical meaning that abandons all reference to the world, in order to avoid the contradictions of earlier notions of set and to achieve a certain conception of conceptual precision. On such terms, from my reality-based perspective, the standard definition of a group has no earthly meaning. But the concept of an abstract group captures something very important and some sort of genus is certainly called for. And, ironically, mathematicians are even making an important distinction in their choice of genus: Not everything studied in mathematics counts as a set and it is critical to distinguish those things that are sets from those that are not.[16]

I discussed, in Chapter 6, the need for a concept of set, a way to rehabilitate the concept, and a proper hierarchy. If one understands sets from this perspective, it might make sense to simply retain the wording of the standard definition, but to reinterpret the word *set* along the lines of Chapter 6. Such a policy, at least in a provisional way, for the sake of following an argument, could be followed generally, whenever one encounters a mathematical definition along the lines of "X is a set such that … ."

However, I believe it would be better to replace the word, *set*, and characterize an abstract group as a system of measurements such as transformations, viewed from an abstract perspective that retains the *distinctions* among the transformations and their *laws of composition*, but treats as omitted measurements everything else about them.

As in the case of vector spaces, there is tremendous value in and need for the abstract-group level of abstraction. But it would be a mistake to forget or ignore the fact that groups first arise as transformations in certain contexts and are meaningful because of those contexts, and/or because of any new context to which they might apply.

If hierarchy is, indeed, often ignored or if the need for it is simply unknown, it remains noteworthy that mathematicians point out, consider important, and prove, a theorem that, in fact, every group can be "realized" as a transformation group.[17] That is, any system of elements possessing a multiplication transformation with the appropriate properties can be considered a transformation group, with the particular object being transformed omitted. I will return to this point later in the chapter.

There is an analogy for the real number system. Numbers are always used to count something, but there is no limit to the sorts of things that they can be used to count: the unit being counted depends on the application. One can study relationships between numbers without having to choose a physical unit. These relationships between numbers will always hold no matter what the unit may be.

In just this way, a particular group will apply to any object or situation that possesses the structure of symmetry relationships captured by the particular group. But the relationships between the elements of the group are independent of any particular object that they might transform. Those relationships between the transformations will always be the same no matter what object the elements might be used to transform. In sum, a group is to the objects it transforms as the number system is to the unit being counted. A group can be considered in abstraction from a particular object that it can transform; the number system can be considered in abstraction from a particular unit that it can be used to count.

However, the concept of an *abstract group* (though usually just referred to as a *group*) is aptly named because it represents a higher level of abstraction than numbers do. The analogy of groups to the number domain is closer when one regards positive numbers in their role as multiplicative *transformations* because a transformation specifies the relationship between the similar things that it relates. Indeed, positive numbers, as we have seen, can be regarded, multiplicatively, as a transformation group acting on magnitudes.

But an *abstract* group omits precisely the specific *relationships* that are measured by its transformations and retains only the distinctions among and the group-multiplicative relationships between the transformations in the group. In contrast, the number system does *not* omit the relationship involved because that relationship is always the same, namely, that of *ratio*, ultimately a relationship to a unit.

Moreover, two transformation groups can differ in structure, whereas there is but one number system. So one needs to combine, as exemplifying the same group, any two transformation groups that have the same structure; one also needs to distinguish any two transformation groups with different structures. Two transformation groups are instances of the same abstract group precisely when they have the same structure. They exemplify the same group when their elements can be matched up in such a way that they have the same times table. This entire question is moot in regards to number.

To classify is to identify, delimit, and conceptualize a range of possibilities. For example, as we saw in Chapter 7, one classifies finite dimensional vector spaces by identifying their respective dimensions. Any two vector spaces of the same dimension have the same structure. One classifies abstract groups, as well, but that classification is exceedingly difficult.[18]

I have said that groups arise as transformation groups, just as numbers are used to count. Groups are meaningful because they can be used to transform or measure symmetries and numbers are meaningful because there are things to count.

But this is not the way groups are typically presented in standard textbooks intended for mathematics majors. Mathematicians always give examples and they know that this is necessary. But they seldom use these examples to *motivate* their concepts. To the extent that such motivation is lacking, groups are presented as if they were a free invention of the human intellect, an invention that just *happens* to have applications to something out there in the real world.[19]

Such an approach, to the extent that it is followed, is both a-historical and anti-conceptual. Historically groups arose precisely in the way that I have indicated, as permutations or transformations reflecting existing symmetries in their objects.

Conceptually, groups have meaning only insofar as they pertain, directly or indirectly, to the world. Having traced that connection, it is clear that groups do pertain to existing relationships in the world. Groups measure relationships involving symmetry. The study of groups, though more specialized than the study of number, is as much the study of quantity, of measurement, as is the study of numbers.

The Structure of Groups

There is a rich variety of distinguishable abstract groups; we have discussed just a few. One unique number system suffices to measure an endless variety of multitudes and magnitudes. By contrast, transformation groups are multi-faceted and are distinguished by their structures. If groups are the means of studying symmetry, one also needs a way to study groups. One needs to get at their structures and to identify the respects in which two groups can differ. As

mathematicians would put it, one needs a *classification* of possible abstract groups. And the first pair of related questions required even to begin such a question are:

1. When are two groups essentially the same abstract group?
2. How are abstract groups distinguished?

I have discussed, in detail, the group of symmetries of an equilateral triangle, the puzzle piece transformation group. Then I introduced a second group, the permutation group for three objects, also known as the full symmetric group on three objects and, in the standard notation, designated S_3.

We saw that, in essence, the two groups have the same the structure. Their respective elements can be lined up and there is an exact correspondence between the transformations of the triangle by the first and the permutations of the three marbles by the second. And, once corresponding elements are identified, they share the same times table. And this very circumstance led to the concept of an abstract group.

The *concept* of an abstract group is an essential first step in addressing this complexity. Everything essential about an abstract group, from a structural perspective, is captured in its times table. If the elements of two groups are lined up in such a way that their times tables are the same, then their structures are identical, no matter what context they may have arisen in and what kind of symmetric object they are utilized to measure. If this cannot be done, then the groups have essentially different structures and measure a different kind of symmetry.

Now a times table gets very big, very fast. A group, containing, say, 100 elements has 10,000 products. This quickly becomes unmanageable. One needs more efficient ways to check that two groups share the same times table.

In this section, I will describe one way that a particular group can be fully *characterized* without having to construct a times table. This characterization will provide a first indication of a very large enterprise.

I have already mentioned that every element of the puzzle piece group is either a rotation or a reflection followed by a rotation. Let us develop this point. To this end, I define $T = C_r$. (Recall that C_r is the transformation that leaves C fixed and interchanges A and B.) For rotations, I use the specific rotation that is already designated by the letter R.

I have chosen the letter T to stand for “transpose”. The letter R (for “rotation”) is already taken and transpose refers to the fact that T transposes the locations of the vertices at puzzle-block locations A and B. Notice also how R and T relate to

the S_3 group. R permutes the marbles by moving the third marble to the beginning. T transposes the first two marbles.
I claim that every element of the puzzle piece group is either a power of R (R multiplied by itself some number of times) or a product of T with a power of R. One can see this directly by verifying, from the times table, the relationships depicted in Figure
16:

E	=	RRR	=	R^3
R	=	R	=	R
R^2	=	RR	=	R^2
A_r	=	RRT	=	R^2T
B_r	=	RT	=	RT
C_r	=	T	=	T

Figure 16

For example, using the times table, one finds, RRT = R(RT) = R(RC_r) = RB_r = A_r. In the rightmost column, the designation of R^2 and R^3 is defined as R multiplied by itself two or, respectively, three times, or, as one usually puts it, R raised to the second or third power.

So every element of the group can be generated from just two of its elements, R and T. As mathematicians put it, R and T are *generators* of the group.[20]

R and T generate the group, so every string of multiplications involving R and T, such as RRRTTTRTTTTRT, is an element of the group. However, there are only six elements in the group so it must be possible to simplify every such combination of multiplications to discover which element is involved. How is this done?

First, recall that RRR = E and TT = E or, alternatively, R^3 = E and T^2 = E. Any string of three R letters can be replaced by the identity, E, and any string of two T letters can also be replaced by the identity. Recall also that EA = AE = A for any element of the group. Applying these rules to the string above, one finds: RRRTTTRTTTTRT = ETTTRTTTTRT = TTTRTTTTRT = ETRTTTTRT = TRTTTTRT = TRETTRT = TRTTRT = TRERT = TRRT.

To simplify this expression further, one needs a way to interchange T and R. We already know that T (= C_r) doesn't commute with R. However, one does have TR = C_rR (by substitution) = A_r (by the times table displayed in Figure 9)= R^2T (by the table displayed above as Figure 15). So TR = R^2T.

From this relationship, remembering that T^2 = E, one also finds, first, that R = T(TR) = TR^2T, simply by multiplying the equation on the left by T. Then one multiplies the resulting equation, R = TR^2T, by T on the right, to obtain RT = TR^2T^2 = TR^2. The equation of interest is between the first and last terms in this string of equalities, namely RT = TR^2. Between this equation and the earlier equation TR = R^2T, one knows the effect of interchanging the order of the transformations R and T, moving T to the right.

These formulae provide a way to simplify the expression to one of the expressions in the above table. One finds TRRT = (TR)RT = (R^2T)RT = R^2TRT = R^2(TR)T = R^2(R^2T)T = R^2R^2TT = RTT (by the times table) = RE = R. Therefore, RRRTTTRTTTTRT = R. By this method, any string of multiplications of R and T can be reduced to one of the six elements in the table.

I have used the following relationships: R^3 = E, T^2 = E, and TR = R^2T. These three relationships completely determine the structure of the puzzle-piece transformation group, also known as the group of symmetries of the equilateral triangle, also known as S_3: the full symmetric group on three objects. As a mathematician would put it, S_3 can be generated by two generators R and T subject to three relations R^3 = E, T^2 = E, and TR = R^2T (where E is understood to be the identity element of the group).[21] Three rules for combining the two generating symbols thus replace a six by six multiplication table and are sufficient to derive the entire times table if that were ever needed.

This is not the only available technique to understanding the structure of a group, but it is an important one. And it illustrates the kind of approach required to simplify and make this task manageable.

Group Representations

To concretize an abstraction is to consider the ways that it applies to concretes,

to consider a range of representative or important referents of the abstraction. Concretization is one of the ways that one ties one's concepts to the world. Concretization helps answer two key questions about any particular abstraction: Namely, "What does it mean?" and "Why does it matter?"

Concretization is at least as important in mathematics as elsewhere. And it is especially important in understanding abstract groups. Indeed, there is a very complex specialty devoted to this pursuit. It is called the theory of *group representations*.

In mathematics, groups arise as transformation groups acting on a geometric object or acting on a mathematical domain consisting of quantities or measurements. But, qua abstract group, the *particular* geometric object, system of quantities, or system of measurements that it may act on is treated as an omitted measurement. The relationships between the elements of the group do not depend upon on any *particular* object on which it acts. Leaving this detail aside, one retains only the distinctness of the particular transformations and the laws of their composition.

A full understanding of transformation groups, as a *category* of mathematical domains, involves two related pursuits. First, a categorical investigation requires a grasp of the full potential range of essentially distinct abstract groups. This is the classification problem. On the other hand, part of understanding any specific abstract group is to understand, in some terms, the potential *range of objects to which it can be applied* and the differences in those applications. One investigates and classifies both the systems of measurements, of a particular kind, and the kinds of quantities to which each can apply.

The second of these investigations feeds into the first. To understand an individual group is, among other things, to be able to compare it with other groups.

I mentioned the classification problem for finite dimensional vector spaces in Chapter 7. We found that the structural difference between two vector spaces reduces to a single characteristic, namely their respective dimensions.

The case of finite abstract groups is far more complex. And part of that problem, in regards to both the abstract groups and the quantities that they measure, is to decide what needs to be distinguished and what kind of variations should be regarded as unimportant.

Understanding the ways that specific groups can act, as transformation groups, is one key to their classification. Consider, for example, the group of transformations of an equilateral triangle in reference to the geometric shape depicted in Figure 17:

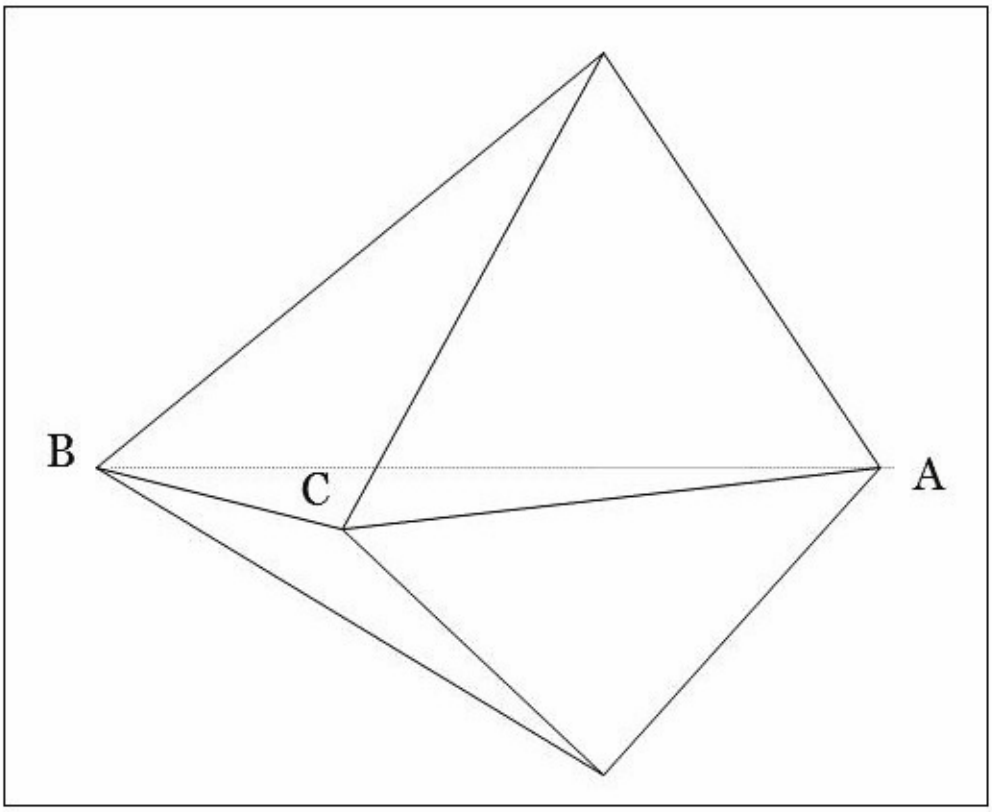


Figure 17

This is a double tetrahedron; the two tetrahedrons are divided by an equilateral triangle.

Suppose that one transforms the triangle with vertices A, B, and C, by moving it, as in the puzzle piece example. When one moves the triangle, the entire double tetrahedron moves with it. More specifically, one moves the figure in such a way that, in the end, the triangle, ABC, occupies the same space that it did before. When the transformation is complete, the entire double tetrahedron occupies the same space that it did before, as well.

Indeed, every placepreserving transformation of the triangle is a placepreserving transformation of the double tetrahedron and, conversely, every placepreserving transformation of the double tetrahedron is a placepreserving transformation of

the triangle. A placepreserving transformation of either one is a placepreserving transformation of the other. In short, the symmetries of the triangle are also, precisely, the symmetries of the double tetrahedron.

There is no essential difference between the placepreserving transformations (or symmetries) of the triangle and those of the double tetrahedron. (I omit *reflections* of the double tetrahedron from consideration.) Yet, by the same token, on a more concrete level, symmetries of an equilateral triangle are not limited, in their application, to the triangle: These symmetries also apply, as in this example, to more complex geometric figures such as the double tetrahedron. And it is important, in the appropriate contexts, to do justice to both perspectives: to the perspective from which things are the same and the perspective from which they are different. Abstract reasoning does not mean that one forgets the referents of the abstractions; conversely, remembering the referents does not mean forgetting the scope of their integrating abstractions.

To take a similar example that will be important later, consider the cube roots of 1.

These cube roots of 1 are, specifically, the three solutions to the polynomial equation $x^3 – 1 = 0$. Factoring, one has, $(x - 1)(x^2 + x + 1) = 0$. The root corresponding to the first factor is, obviously, $x = 1$. According to the quadratic formula, the solutions to the second factor, to $x^2 + x + 1 = 0$, are $x = -1/2 + (i/2)\sqrt{3}$ and $-1/2 - (i/2)\sqrt{3}$ where $i = \sqrt{(-1)}$ is the square root of minus 1. (One can verify this directly, as well.)

It will be convenient to use the symbol ω to designate $-1/2 + (i/2)\sqrt{3}$. By simply multiplying, one finds that $\omega^2 = (-1/2 + (i/2)\sqrt{3})^2 = -1/2 - (i/2)\sqrt{3}$ and $\omega^3 = 1$. In other words, ω, ω^2, and ω^3 are all cube roots of 1.[22]

By either direct addition or by simply appealing to the fact that ω is a solution of the equation $x^2 + x + 1 = 0$, one also has $\omega^2 + \omega + 1 = 0$. So the cube roots of 1 sum to zero. For the intrigued: This is a general phenomenon. For any positive integer n, for essentially the same reason, the nth roots of unity sum to zero.

Finally, there is the well known operation of complex conjugation that replaces any complex number, a + bi, with the complex number a – bi.[23] Clearly, the complex conjugate of $-1/2 + (i/2)\sqrt{3}$ is $-1/2 - (i/2)\sqrt{3}$. Or, in my terminology, the complex conjugate of ω is ω^2, and vice versa.

A graphical perspective of complex numbers displays the real numbers as the x-axis and places the complex number I at the point y = 1 on the y-axis. When one uses complex numbers, in this way, to measure the plane, one refers to this

system of measurements as the *complex plane*.
With this geometric application of complex numbers, one finds that both ω and ω^2 lie on the unit circle. To see this, let r be the length of the line drawn from zero (coordinates (0, 0)) to one of the numbers ω or ω^2. By the Pythagorean theorem,
$r^2 = (-½)^2 + (½\sqrt{3})^2 = ¼ + ¾ = 1$.
The value of r given by this formula is known as the *modulus* of the complex number. In general, the modulus of a complex number, a + bi, is $\sqrt{(a^2 + b^2)}$. The modulus of a complex number z = a + bi is traditionally written |z|.[24]
It is well known that the modulus of the product of two complex numbers is equal to the product of their moduli.[25] To see this directly, calculate:

$|(a + bi)(c + di)|^2 = |(ac - bd) + (ad + bc)i)|^2 = (ac - bd)^2 + (ad + bc)^2 = (a^2 + b^2)(c^2 + d^2) = |a + bi|^2|c + di|^2$

Geometrically, one represents this situation as in Figure 18:

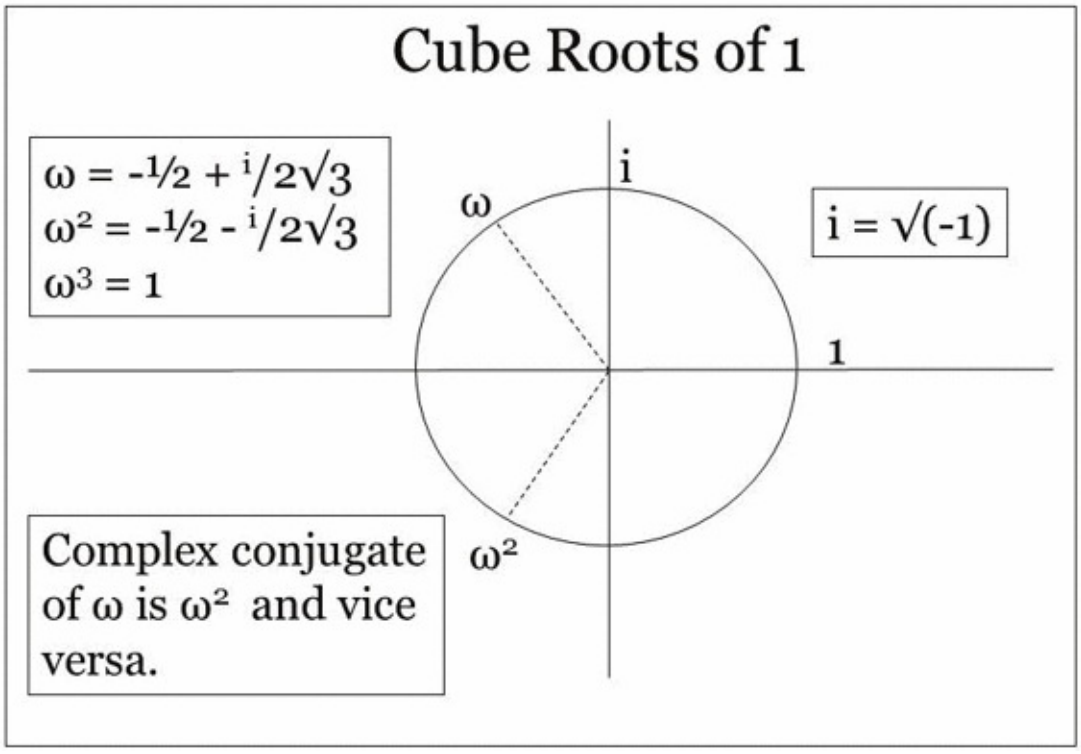


Figure 18

It is well known that multiplication by a complex number of modulus 1 has the effect of rotating the complex plane. The direction of that rotation is counter-clockwise and the magnitude of the rotation is the angle, with respect to the positive real-axis of the line drawn from zero to the complex number by which

one is multiplying.[26]

In particular, the effect of multiplication by ω is to rotate each of the cube roots of unity, along with everything else, by 120^0. For one has $\omega \times 1 = \omega$, $\omega \times \omega = \omega^2$, and $\omega \times \omega^2 = 1$. Every multiplication by ω rotates the complex plane and the third such multiplication brings the cubes roots of unity, along with everything else, back to its original position. Since ω^3 represents a rotation by 360^0, multiplication by ω is one third of that, namely, 120^0.

Notice also that the effect of complex conjugation is to leave the number 1 fixed and, more broadly, to reflect everything in the complex plane with respect to the real axis, thus interchanging ω and ω^2.

In sum, multiplication byωactstopermutethecuberoots of 1, as does complex conjugation. The first is a rotation and the second is a transposition. Together, the rotation and the transposition generate a group of permutations of the cube roots of unity.

Notice that the circle is transformed as well. The rotations rotate the circle, while complex conjugation reflects it about the real (horizontal) axis. Together, multiplication by ω and complex conjugation generate a group of transformations of the circle. It is, specifically, a group that measures a certain kind of symmetry embodied in the above diagram. To wit, it is a group of transformations of the circle onto itself that *preserves the set* of points $\{1, \omega, \omega^2\}$. The transformation group preserves the place occupied by the circle, as well as the place of a particular set of points lying on that circle, while moving the entire circle as a whole.

Furthermore, every one of these transformations constitutes a permutation of elements in the set $\{1, \omega, \omega^2\}$. Since these two operations, namely, multiplication by ω, and complex conjugation, are sufficient to generate all six permutations of $\{1, \omega, \omega^2\}$, this group of transformations is none other than S_3. One can see this, finally, by directly relating both these symmetries, and their generators, to the symmetries of the equilateral triangle.

Imagine, then, that the A corner of the equilateral triangle is at ω, the B corner is at ω^2, and the C corner is at 1. Then multiplication by ω has the same effect as R and complex conjugation has the same effect as T. This situation is captured in

Figure 19:

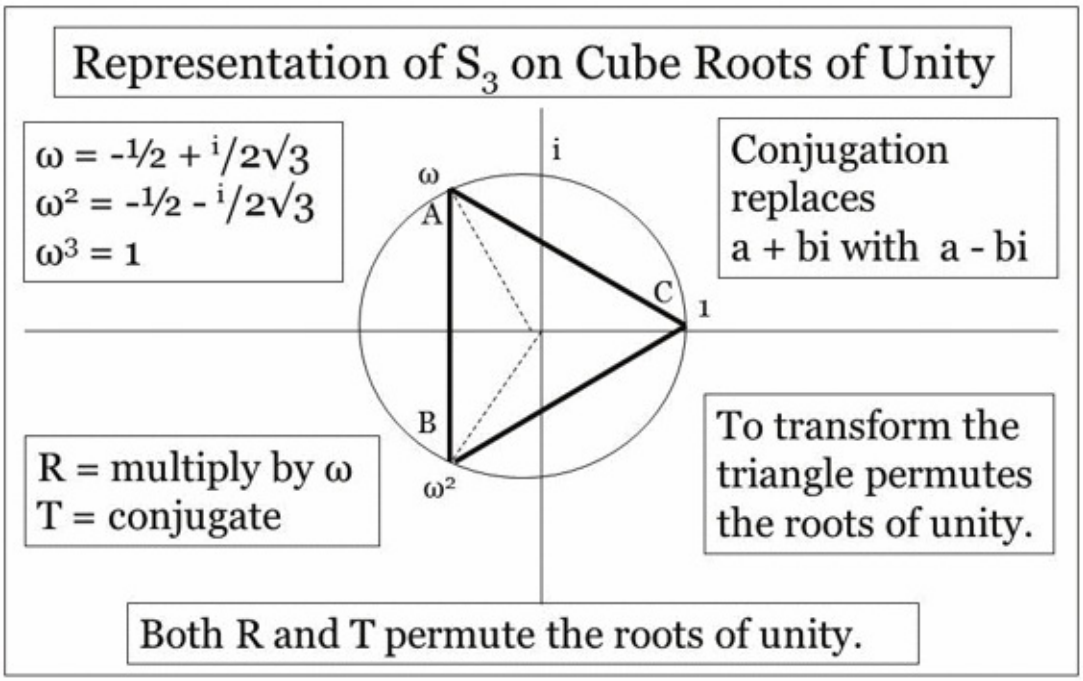


Figure 19

Geometrically, the two situations are identical. But the use of complex numbers moves that example to an algebraic context, capturing the geometric symmetries in an analytical form.

So far I've applied S_3 to figures whose symmetries are completely exhausted by symmetries in S_3. But it's also possible for S_3 to act on more complex figures containing yet further symmetries. For example, Figure 20 depicts a slightly more complex puzzle block than my original example:

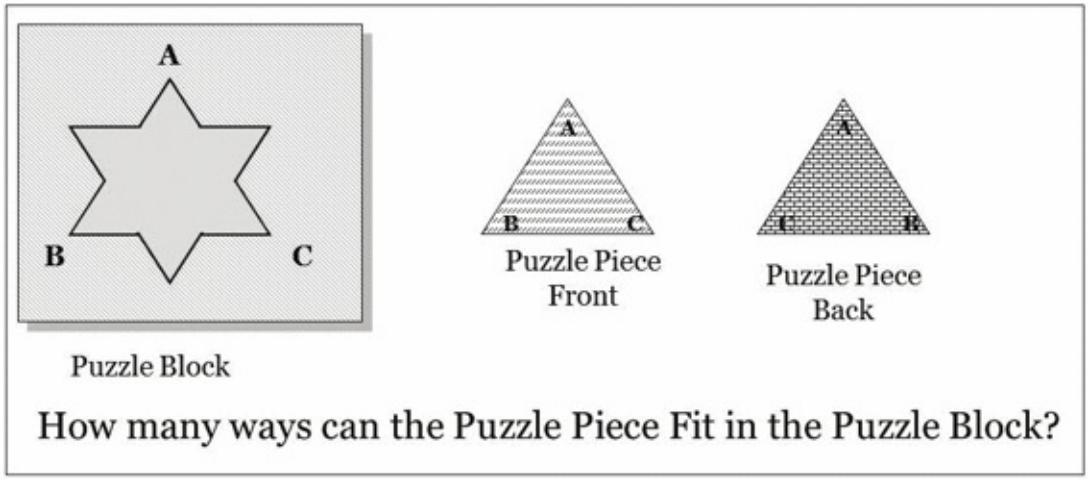


Figure 20

In this case, every solution to the first puzzle is also a solution to this one. But, as indicated in Figure 21, there are an additional six solutions, as well, three of them showing the green side and three of them showing the pink:

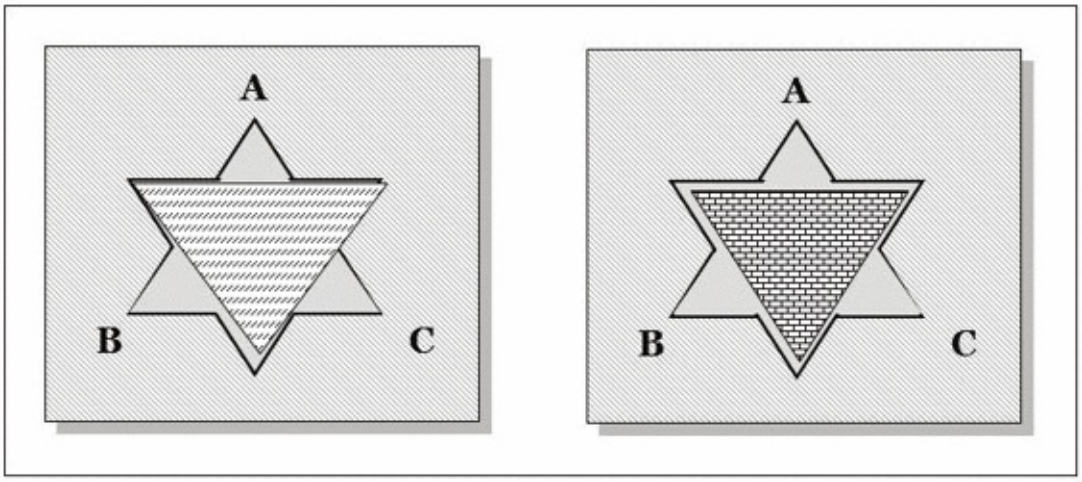


Figure 21

In all, there are twelve different permutations and the full transformation group has twelve transformations. The original set of permutations, S_3, is a subset of the full set of 12 and S_3 is, therefore, a subgroup of the larger group.

Yet another example has a much wider significance. I return to the group of transformations, S_3 that I started with. To make this discussion easier to follow, I use the generators of S_3, namely R (counter-clockwise rotation) and T (transposition of A and B). In the last section, I established the defining relations of these generators: $R^3 = E$, $T^2 = E$, and $TR = R^2T$ (where E is understood to be the identity element of the group). In terms of these generators, the

distinguishable elements of the group are E, R, R^2, T, RT, and R^2T.

Consider a set of six marbles, each one labeled, in Figure 22, with one of these designations:

One can regard these labels as coded instructions for permuting the marbles. Assume that the marbles are arranged in a row or column. Any element of S_3 determines a permutation in the following manner. Multiply the transformation named on each marble, on the left, by the chosen element of S_3. Then, replace the marble by the marble labeled with the resulting element of S_3. The result is a permutation of the marbles.

For example, applying the element R in S_3 yields the permutation in Figure 23:

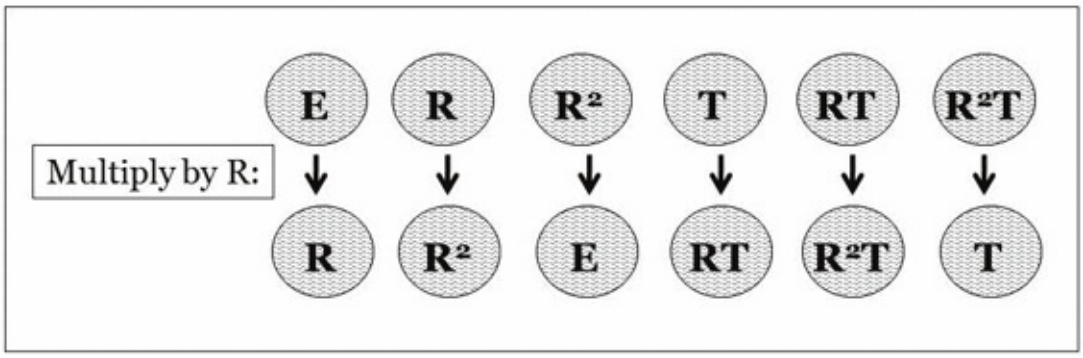


Figure 23

Since, for example, multiplying T on the left by R results in the group element RT, the marble with the label of T is replaced by the marble with the label of RT.

Every such multiplication is reversible, because every element in a group is invertible. Therefore, the mapping is one-toone. Indeed, if a, b, and c are elements of S_3, and b and c are different elements, one cannot have ab = ac. For, if this equation held, multiplying the equation on the left by a^{-1} would yield b = c, contrary to assumption. Consequently, applying any element of S_3 in this fashion yields a permutation of the marbles.

As a second example, Figure 24 depicts the effect of applying the transposition element T:
Since the elements R and T generate the entire transformation group S_3, every transformation by elements of S_3 of the set of marbles can be resolved into a series of transformations by R and T. So the effects of T and R, as depicted

together in Figure 25, are enough to determine this set of permutations:

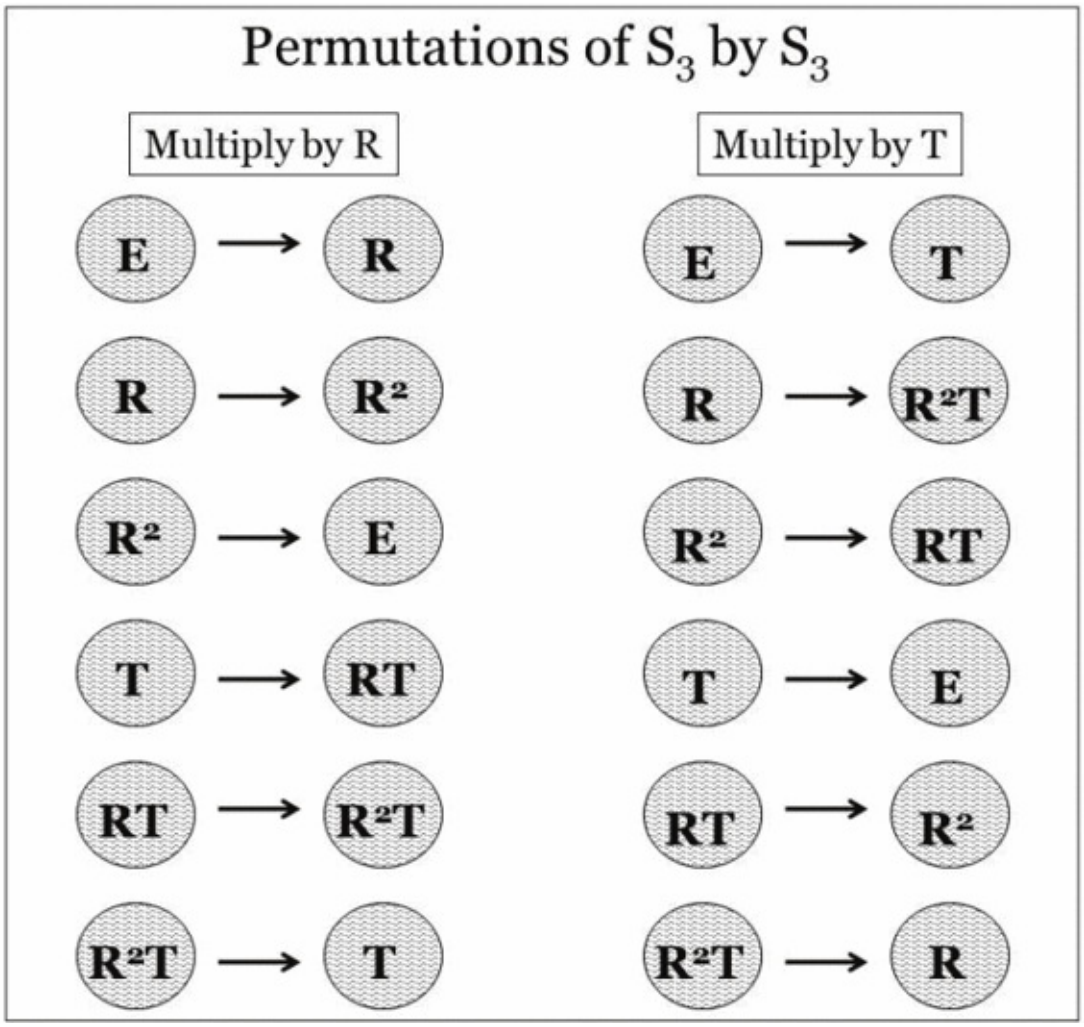


Figure 25

I have labeled this picture, "Permutations of S_3 by S_3," for a reason: Nothing in this discussion depends upon the particular objects to which these labels were applied. One could, by the same token permuted boxes, spoons, or any other set of six objects that one chose to distinguish by this set of labels.

In particular, one can apply these permutations directly to the elements of S_3. Starting with this observation, there are a number of points worth noting about this example:

● The six elements of S_3 account for six permutations of S_3.

This is a rather small subset of the total number of permutations of six objects: Recall that the total number of permutations of six objects is 6! (=6×5×4×3×2×1 = 720).

● Nothing about this discussion required knowing anything at all about S_3 beyond its set of elements and the rules for multiplying them. The argument applies to any abstract group for which the elements are specified and for which either a set of relations (as in this case) is given or its times-table is given. How the group multiplication is specified does not matter. It suffices that one has specified that multiplication for any two elements of the group.

● In conclusion, any abstract group can be realized as a permutation group acting on its elements, or acting on any set of objects, such as the marbles in my example, that have been distinguished by the elements of the group.

● Or, more simply, any abstract group can be realized as a transformation group of permutations. This is known as Cayley's Representation Theorem.[27]

● No two different elements of the group produce the same permutation. One says, therefore, that Cayley's representation is a *faithful* representation.

Generally speaking, in considering a mathematical alternative, one may not know, in advance, whether any specific potential alternative will ever be realized. In regards to a particular number, for example, one does not know that it will ever be needed, but, in defining the number system as a system of measures, one has provided for the eventuality. And it is the job of mathematics to provide, in advance, for such eventualities so that, when they are encountered, one will already know how to deal with them. Keep in mind the presumption that one has identified what numbers are, has identified, in general, why numbers are important, and has identified the way that numbers relate to each other. It is for this reason that one is able to provide, and properly motivated to provide, for the entire range of numbers in advance of a specific identified need for most of the particular numbers in the system of measurements.

Similarly, in identifying, in some form, a particular abstract group, one may not know, in advance, whether it will ever be needed. But it is valuable to anticipate the possibility. First, one knows, generally, that groups of transformations are important. This implies, by Cayley's Representation Theorem, that abstract groups are, in general, important. One knows that any particular abstract group belongs in the list of reasonable possibilities, worthy of general consideration. Second, granting the general importance of finite groups, one knows, from this discussion, how any particular finite group would relate to concretes. For in understanding finite groups as permutations, one also has a starting point to

understanding finite groups as permutations, one also has a starting point to investigate each finite abstract group as a general category of systems of measurement.

In this sense, when one considers a particular finite abstract group, one can take its potential applicability to the world, and its potential importance, as a given.

Understanding that any abstract group can act as a group of permutations, is one first, small step on the road to classifying finite groups and their representations.

Matrix Representations of Finite Groups

The standard theory of group representations, however, takes a somewhat different, and marvelously productive, turn. The mathematical theory is quite complex and well beyond the scope of this book. However, its conceptual underpinnings are illuminating and more widely accessible.

Once again, the group S_3 can serve as an example. Consider, as its object, the vector space R^3. (Because I am using the letter R to designate a rotation, I distinguish this expression from my earlier use of that letter by designating the real number line as R.) Suppose that one has chosen a coordinate system for R^3, that one has chosen an x axis, a y axis, and a z axis.

These are *three* axes and S_3 is the permutation group of three elements. So, one can use S_3 to study the permutations of the x, y, and z axes. Applying these permutations, assume, as well, that these permutations carry the rest of the vector space along with them, just like the example of the double tetrahedron. Then the six permutations of S_3 are all represented by matrices acting on R^3.

In particular, assume that the element T acts on R^3 by interchanging the x and y axis and that the transformation R acts on R^3, in a counter-clockwise fashion, by rotating the z axis to the x axis, the y axis to the z axis, and the x axis to the y axis Then, for T, one has

$$T = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

because

$$\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$$

$$\text{and} \quad \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

The first two of these equations imply that the matrix T interchanges the x axis and the y axis. The third equation implies that the matrix T preserves the z axis. Together they imply that, as applied to any vector, the matrix T will interchange the first two coordinates and preserve the third. In other words:

$$\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} b \\ a \\ c \end{pmatrix}$$ To permute the *axes* of R^3 is, from this perspective, to permute the *coordinates* of vectors in R^3.

Secondly, for R, one has

$$\mathbf{R} = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$ because

$$\begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

$$\text{and} \quad \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$$

These relationships can be verified by checking the matrix multiplication. They imply, in turn, that the matrix R rotates the x axis to the y axis, the y axis to the z axis, and the z axis to the x axis. Consequently, the matrix R will rotate the coordinates of any vector to which it is applied. In other words:

$$\begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}\begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} c \\ a \\ b \end{pmatrix}$$

Once again, to rotate the *axes* of R^3 is to permute the *coordinates* of any vector in R^3. It follows that any action on the vector space R^3, of any product of the actions of R and T, is, at the very same time, a permutation of the coordinates of vectors in R^3.

One can, indeed, start from this perspective. I speak generally of the vector spaces R^n: Any transformation of vectors in R^n that acts to permute the coordinates of its vectors is a linear transformation that can, therefore, be represented by a matrix. To specify a permutation on the coordinates of a vector is to specify a matrix that acts on the vector space, a matrix that acts, indeed, to permute the coordinates of the vectors in the vector space.

Consider, in particular, the current example. One finds, by performing the relevant matrix multiplications, that one can establish the entire set of matrices representing elements of S_3, as shown in Figure 26:

$$E = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad R = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \quad T = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$R^2 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} \quad RT = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix} \quad R^2T = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

Figure 26

I have shown that every finite group can be represented as a permutation group and, indeed, that the so-called Cayley representation is a faithful representation for which each group element determines a different permutation than any other group element. Consequently, in just this fashion, for a sufficiently high value of n, one can also represent any finite group as matrices permuting the coordinates of vectors in R^n. In fact, since the Cayley representation is a representation of the

group on itself, the value of n corresponding to, and guaranteed by, the Cayley representation is the number of elements in the group.

It is very important to notice a number of things about this S_3 example:

1. First, there is a complete one-to-one correspondence between the elements of S_3 and the associated matrices that represent its action on R^3.

2. Each of these matrices is invertible.
3. Multiplication of these matrices corresponds exactly to multiplication of elements of S_3.
4. In particular, the identity matrix corresponds to the identity element of the group S_3.
5. As a consequence of the third point, the inverse of each of the matrices corresponds to the inverse of the corresponding element of S_3. For example, the inverse of the matrix corresponding to R is the matrix corresponding to R^{-1}. Since the inverse of R is R^2, that matrix corresponding to R^{-1} is the matrix corresponding to group element R^2.
6. A matrix or a linear transformation on a vector space is called an automorphism if it transforms the vector space into itself and is invertible. All of the matrices corresponding to the action of the Group S_3 on the vector space R^3 are automorphisms of the vector space R^3.

An action of a group G, as automorphisms of, or invertible matrices relating to, a vector space V is called a *group representation* if it satisfies conditions 3 and 4 (and, therefore, 2 – 6). More specifically, the action is called a *representation of G on the vector space*.

A representation does not have to satisfy the first condition. A group representation is still a representation if two group elements act in the same way. For example, as an extreme case, if one associates every element of a group G to an n × n (read: ‘n by n’, in this context) identity matrix, the result is a representation of G, albeit a trivial representation. For the product of the two identity matrices, corresponding to two group elements, is the identity matrix. And this identity matrix, by assumption, is the matrix that corresponds to the product of those group elements.

Matrix representations of finite groups have one particularly striking property.

To begin with, let G by any finite group and let g be an element of G. As one takes successive powers of g, namely, g, g^2, g^3, …, one must ultimately find two group elements in this series that are equal because, after all, G has a finite number of elements and so there are only a finite number of possible values in this series. Therefore, there exist distinct numbers n and m, such that $g^n = g^m$.

It follows from this equation that there is a positive integer q, namely the difference between n and m, for which $g^q = e$. If for, example, $n > m$, one sees this simply by dividing both sides by g^m. If p is the smallest positive integer for which $g^p = e$, one says that the group element g is *of order p*.

This has an interesting, and far-reaching, consequence to group representations. Suppose given, a matrix representation of G. Suppose that an n by n matrix, M, corresponds, in this representation to the element, g. In the context of a matrix representation, the powers of M correspond to powers of the group element g that it represents. In other words, for any positive integer n, the matrix M^n corresponds to the group element g^n. In particular if g is of order q, then the matrix M must satisfy $M^q = I$, where I is the identity matrix.

One can put it this way: In the domain of n by n matrices, M is a *q^{th} root of the identity matrix I*. Clearly, this is true in general. Any matrix, A, representing a specific group element, from any finite group, acting on a vector space, will be a root of unity. That is to say, for any such A, there exists a positive integer q such that $A^q = I$. Finally, for those familiar with determinants, I want to mention an important implication: that the *determinant* of A (a number) is an nth root of unity, for some positive integer n, in the usual sense.

For this reason, one should expect that roots of unity will play an important role in the theory of group representations. And, further, since complex numbers are essential to the study of roots of unity, complex numbers also play a key role in the theory of group representations.

Irreducible Representations

Classifying group representations requires a building-block approach. If one can identify an appropriate class of building blocks and determine the way that these building blocks fit together, one has achieved an important grasp of the range of possible group representations.

possible group representations.

As a close analogy, when physicists search for the elementary constituents of matter, e.g., elementary particles, and study the ways that they interact and combine, they are finding and studying building blocks and the way that these building blocks fit together.

And, as it happens, one of their tools is the mathematical theory of group representations. Nor should this be surprising to those who understand the importance of fundamental symmetries in physical science.

In general, a building block should possess some kind of irreducibility. To analyze a complex structure, one breaks it, in some particular respect, into pieces. As one proceeds in this fashion, one ultimately finds that further analysis, *in this particular respect*, can be carried no further. From the perspective of this particular mode of analysis, one has found the building blocks. Further analysis requires a separate study, from a different perspective, of the building blocks.

In the case of group representations, these building blocks are called *irreducible representations*. As I will indicate, an irreducible representation captures, in a particularly revealing way, an aspect of a group's permutation and geometric symmetries.

So what is an irreducible representation: irreducible as opposed to what? To understand *irreducible representations*, one must first identify what a *reduction* would consist of. To that subject I turn.[28]

To understand what a reduction consists of, it is helpful to first understand how building blocks may be put together. So, I begin by proceeding in the opposite direction, by extending *representations* on a vector space V to representations on larger vector spaces containing the vector space V as a subspace.

In the last section, I presented a representation of the symmetry group S_3 on R^3 and I will further develop this example. Throughout this discussion, I will look at that action as a permutation of the *coordinates* of R^3. And I will also characterize extensions to other vector spaces in terms of their effects on coordinates. To recall, the S_3 element R acts on vectors in R^3 by rotating their coordinates, moving the third coordinate to the first position, and T acts by interchanging the first two coordinates.

Suppose that one wants to extend the representation of S_3 to a larger vector space that contains the vector space R^3. To extend a representation is to find a representation on the larger space that restricts to the original representation when restricted to the original smaller subspace. There are some straightforward ways of doing so.

Consider R^4 as a first example. S_3 acts on the first three coordinates of $R^4$by permuting them, as in the action on R^3. What is the simplest way to extend this action to R^4? Simply leave the fourth coordinate alone! In this way, every action on the first three coordinates is automatically extended to an action on, to a permutation of, all four coordinates.

To see this explicitly, the transformation T acts on vectors in R^3 as follows: T(a, b, c) = (b, a, c). To extend that action to R^4, set T(a, b, c, d) = (b, a, c, d). Similarly, extend R(a, b, c) = (c, a, b) to R^4 by setting T(a, b, c, d) = (c, a, b, d). The actions of other elements of S_3 on R^4 are generated by taking products of R and T.

Notice that this particular representation on R^4, by explicit intention, leaves the fourth coordinate untouched. If follows that the vector subspace, W, consisting of vectors (a, b, c, 0), is left invariant by this particular action of S_3 on the vector space R^4. The action of S_3 on any vector in W is a vector in W. To put it another way, this action of S_3 on the vector space R^4 *reduces* to an action of S_3 on a *subspace* of the vector space R^4, namely the subspace W consisting of vectors of the form (a, b, c, 0).

We have extended the action S_3 to an action on R^4, but that very action can, in turn, be reduced to an action on a proper subset of R^4, namely the subspace from which it was extended in the first place.

As in the case of R^3, one can express the action on R^4 by 4×4 matrices (read: ‘4 by 4 matrices’). These matrices permute the coordinates of R^4. For example, to permute the first two coordinates in R^3 (the action of T) *is*, when extended to R^4, to permute the first two coordinates of any vector in R^4. And to permute coordinates is to permute the basis vectors corresponding to those coordinates.

Finally, notice that the subspace V consisting of all vectors of the form v = (0, 0, 0, d) is also, trivially, invariant (transformed into itself) under the action of S_3. That is, for any vector v in V, Tv = v and Rv = v, and, therefore, the action of any element of S_3 on v is to multiply it by 1. This is important because every vector in R^4 can be expressed uniquely as a sum of vectors of the form (a, b, c, 0) + (0, 0, 0, d), that is as a sum of vectors from these two invariant subspaces of R^4. In this sense, R^4 is a sum of two subspaces, each of them invariant under the action of the representation. One says that the subspaces decompose the representation.

To *qualify* as a reduction of a representation, it is enough to find an invariant subspace. But it turns out, and one proves, that such a decomposition of a reducible representation is always possible.

One need not stop at R^4! One can extend the action of S_3on R^3, in exactly the same way, to an action on R^n. Simply leave every coordinate in R^n, after the first three, in place under every action by elements of S_3.

Such an action is reducible in the same way that the extension to R^4 was reducible. Specifically, S_3 acts on the subspace W of R^n, for which all coordinates after the first three coordinates are zero simply by permuting the first three coordinates in the prescribed manner. And the reason that I am able to say that S_3 *acts on that subspace* is that the action of an element of S_3 on any vector in W is a vector in W. Again, one says that the subspace W is *invariant* (transformed into itself) by the action of S_3 on R^n.

Once again, one decomposes R^n into two invariant subspaces, into one for which *all but* the first three coordinates are zero and one for which the first three coordinates are zero. The representation acts as the identity matrix on the second of these subspaces.

There is a somewhat more interesting extension to R^6. Apply every permutation in S_3, simultaneously, to both the first three coordinates and the last three coordinates of any vector in R^6.

For example, the action of T on R^3 is $T(x_1, x_2, x_3) = (x_2, x_1, x_3)$. The extension

to R^6 is provided by the formula $T(x_1, x_2, x_3, x_4, x_5, x_6) = (x_2, x_1, x_3, x_5, x_4, x_6)$.

Once again, this action can be reduced to an action on a subspace of R^6, namely to the subspace W of vectors of the form, $(x_1, x_2, x_3, 0, 0, 0)$. For example, T acts on this subspace by the formula $T(x_1, x_2, x_3, 0, 0, 0) = (x_2, x_1, x_3, 0, 0, 0)$. While it is true that the action of T on other vectors in R^6 is more interesting than this, it remains true that the subspace W is invariant under this action of S_3 on R^6. And, once again, the invariance of the action on W is inherent in the fact that this particular action on R^6 was designed to extend an action on W.

And, of course, S_3 acts, in a completely analogous fashion, on another invariant subspace of R^6, namely the subspace for which the first three coordinates are zero. Once again, R^6 decomposes into a sum of two invariant subspaces.

Not all actions of groups on vector spaces permute their coordinates. But consideration of those actions that do is enough to indicate the general pattern of the way that representations of a group on vector spaces can be combined to specify still more complex representations of larger vector spaces.

I have indicated how S_3 can act to permute the coordinates of R^3. It can also act to permute the coordinates of R^2.
I define such a representation, as follows: Think of the permutations of the coordinates of R^2 as corresponding to states of the triangle puzzle. In one state the green side is showing. This state corresponds to the starting positions of the x and y coordinates of vectors in R^2. In the other state of the triangle, the red side is showing. This side corresponds to permuting the x and y coordinates of vectors in R^2. Define the action of T on R^2 by requiring that it switch the x and y coordinates turning the vector expression, in effect, upside down.
Next, just as any pure rotation in S_3 leaves the green side of the puzzle piece showing, so it should also leave the x and y coordinates unchanged. So define the action of R on R^2 to be the trivial action, namely the identity transformation. Between them, these actions of T and R determine an action of S_3 on R^2.
In this action, any of the transpositions T, RT, or R^2T will switch the x and y coordinates. In short, any of the group elements E, R, or R^2, act as the identity matrix, while T, RT, and R^2T all act as the matrix:

$$T = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

In this case, certain distinct elements of S_3, for example, RT and R^2T, have the same action on R^2. Each transposition permutes the two coordinates. This is the first non-trivial example of this phenomenon that we have seen.

This example is no coincidence. It is not an accident that this action is, in effect, an action of the quotient group that I discussed earlier in the chapter. The action of pure rotations, in this representation, is trivial, corresponding to the element e of the quotient group. What remains is the action of T, corresponding to the element r of the quotient group. This action of S_3 on R^2depends only on the equivalence class of each element. Again, any of the transpositions T, RT, or R^2T multiply, in effect, by T; rotations multiply by 1.

This leads to another interesting action of S_3 on R^5. It suffices to specify that action for R and T. In this representation, let R act to rotate the first three coordinates, leaving the last two alone. Next, let T switch the first two coordinates (as its action on R^3) and, also, the last two coordinates (as its action on R^2). These actions correspond to the following 5×5 matrices:

$$R = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \text{ and } T = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

This last example indicates a more general pattern. It combines a permutation on the first set of coordinates with a different permutation of the second set of coordinates.

Once again, this is a reducible representation. It acts, in an invariant fashion, on the subspace W of vectors for which the last two coordinates are zero and, also, on the subspace U for which the first three coordinates are zero.

In sum, in all of these particular examples, the extended actions of S_3 were all reducible to actions on the extended vector spaces for which all of the coordinates after the first three were set equal to zero. A non-trivial extension

can always be reduced to the subset from which it originated.

More generally, a representation of a group G on a vector space V is reducible if there is a proper subspace W for which the action of any element g of G on any vector w within the subspace W results in a vector also contained in W. A proper subspace is a subspace that is not equal to the entire vector space. Symbolically, if g ε G and w ε W then gw ε W. (The juxtaposition, gw, is a symbolic expression of the action of g on w.)

If there exists no such reduction of a representation to a representation on a subspace, then the representation is said to be *irreducible*.[29]

What about the representation of S_3 on R^3 that started this discussion? Is this an irreducible representation? The interesting answer is: No!

Finding irreducible representations is a very complex, though very interesting, undertaking. There are very general techniques for doing so, but these techniques are well beyond the scope of this discussion. So to proceed, I will simply present the outcome for S_3, which can be understood without regard to how one might have discovered that outcome in the first place.

First, keep in mind that S_3 is generated by the rotation R and the transposition T. R acts on any vector in R^3 by rotating its coordinates and T acts on any vector in R^3 by permuting its first two coordinates. Notice the following:

1. Both R and T preserve the sum of the three coordinates. To permute three coordinates does not affect their sum.
2. If all three coordinates of a vector are equal, both R and T transform the vector to itself. If three coordinates are equal, permuting them has no visible effect.

But R and T, together, generate S_3 so these statements apply, as well, to all elements of S_3. Indeed, one sees both points directly. As for the first, all elements of S_3 permute the coordinates of vectors in R^3 and, therefore, preserve the sums of those coordinates.

As for point 2, vectors for which all three coordinates are equal comprise a onedimensional subspace U of R^3. All elements of S_3 act as the identity

transformation on U. Clearly, this is a representation of S_3, namely the trivial representation, on a onedimensional subspace of R^3.

Now, look at vectors v in R^3 for which the *sum* of the coordinates, a, b, and c, is zero.
Such vectors constitute a two-dimensional subspace V of R^3. Notice, first, that the sum of any two vectors in V is a vector for which the sum of the coordinates is zero. It is, therefore, a vector in V. Next, observe that the product of any vector in V by a number is a vector for which the coordinates sum to zero. So it is also a vector in V. More generally, any linear combination of vectors in V is a vector in V, because any linear combination can be generated from a sequence of products and sums of this kind. As to dimensionality, one can freely choose any two coordinates. The third is forced by the requirements that the coordinates sum to zero for all vectors in V. Finally, we have already observed that V is invariant under the action of S_3, because these actions do not affect the sum of the coordinates.
Since S_3 restricts to an action on a subspace of R^3, the action of S_3 on R^3 is reducible.
What about this action on V? Suppose one chooses, as a basis for V, the two vectors

$$\mathbf{v_1} = \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix} \text{ and } \mathbf{v_2} = \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix}$$

One notices a certain level of symmetry in this choice of basis because $Tv_1 = v_2$ and $Tv_2 = v_1$. However, the effect of R on these two basis vectors is not particularly illuminating. One has

$$\mathbf{Rv_1} = \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix} = \mathbf{v_2 - v_1} \text{ and } \mathbf{Rv_2} = \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix} = \mathbf{-v_1}$$

These equations for R and T certainly confirm that V is invariant under the action of elements of S_3, but there is nothing symmetric about this outcome. However, this is the best one can do without using complex numbers. And, it's

worth exploring what complex numbers can add to this picture.

So far, I have not discussed the use of complex numbers as coefficients, or coordinates, of vectors in vector spaces. I have, without explicit discussion, limited my discussion of vector spaces to vector spaces "over the real numbers," that is, to vector spaces with real coefficients and coordinates. But the use of complex numbers is an enormous convenience in mathematics and, in particular, in the theory of irreducible representations.

Primarily, this is for two reasons, one geometric and one analytic. The analytical reason is that any polynomial of degree n, has precisely n roots, providing that one permits complex numbers among those roots and keeps track of multiplicity. If, on the contrary, one insists on restricting one's search to real numbers then equations such as $x^2 + 1 = 1$ have no solutions at all. This is relevant to this discussion because polynomials with complex roots are ubiquitous in the study of matrices and, especially, in the theory of group representation.

Omitting complex numbers would be more than just a technical inconvenience. In general, polynomials and their roots come up constantly in mathematics. And these complex solutions have application. They are meaningful and important. One could possibly find a way to live without *explicitly* using complex numbers, but it would be incredibly inconvenient! And any such attempt would, in the end, involve smuggling in a disguised or reincarnated version of complex numbers. In substance, complex numbers, whatever one might call them or how one might represent them, are unavoidable.

The geometric reason derives from the geometric interpretation of complex numbers that I discussed earlier. Multiplication by $i = \sqrt{(-1)}$, from a geometric perspective, is counter-clockwise rotation by 90^0. Complex numbers provide a convenient and powerful way of treating angles analytically and this is one of the secrets of their interest, their power, and, indeed, their ubiquity in mathematics.

Both factors are at play in the theory of irreducible representations. First, the use of complex numbers provides a geometric perspective on group actions – as we saw earlier in our application of the cube roots of unity. Indeed, I have, since, explained the importance of roots of unity in the theory of group representations. But every non-trivial root of unity, every root of unity other than 1 and -1, are complex numbers. The cube roots of unity, in particular, are the three roots of

the polynomial equation $x^3 - 1 = 0$.

Regarding the action of S_3,recall myearliernotation:$\omega = 1/2 + (i/2)\sqrt{3}$. Consider the action of R and T on the vectors

$$v_1 = \begin{pmatrix} \omega \\ \omega^2 \\ 1 \end{pmatrix} \text{ and } v_2 = \begin{pmatrix} \omega^2 \\ \omega \\ 1 \end{pmatrix}$$

These vectorsbelongtoVbecause,asIpointedoutearlier,$\omega^2 + \omega + 1 = 0$: The sum of their coordinates is zero.
One has, as with the previous basis of V, $Tv_1 = v_2$ and $Tv_2 = v_1$. But one also has

$$Rv_1 = \begin{pmatrix} 1 \\ \omega \\ \omega^2 \end{pmatrix} = \omega^2 v_1 \text{ and } Rv_2 = \begin{pmatrix} 1 \\ \omega^2 \\ \omega \end{pmatrix} = \omega v_2$$

One could not imagine a greater symmetry than this. The transposition matrix, T, switches the two basis vectors by switching their first two coordinates and the rotation matrix, R, rotates the coordinates of each. And this rotation of coordinates, by R, is accomplished analytically in the simplest possible way: one multiplies the vector by a complex number. The vector space that S_3 is acting on is two-dimensional, but the action of S_3 is determined as, and, in effect, generated by, the set of permutations of the coordinates of a single vector, namely, v_1. The action of S_3 on these basis vectors is a transparent reflection of the symmetries inherent in the group S_3.

I want to explore this just a little further. Recall that we have, throughout, viewed the action of S_3 as permuting the coordinatesof R^3. As applied to the vector v_1, in acting on that vector, S_3 also acts to permute the three cube roots of unity. Three of these actions, the rotations (including E), accomplish this permutation by simply multiplying $v_1$by – none other than – a cube root of unity. The other three, those elements that involve a transposition, transform v_1 to a multiple of v_2 and, again, that multiplier is a cube root of unity. For example, $RTv_1 = Rv_2 = \omega v_2$.

Moreover, the actions on v_1 and v_2 are totally symmetric, they work the same way. The *rotations* act on v_1 and v_2 by, respectively, multiplying each by a cube

root of unity. The transpositions transform each basis vector to a multiple of the other: v_2 to a multiple of $v_1$and v_1 to a multiple of v_2.

One cannot do better than this and still distinguish, by their actions, every element of S_3. Two dimensions are required, and are sufficient, to fully capture the symmetries of S_3.

That one can do this well is remarkable. But the simplicity of this picture depends totally on the use of the cube roots of unity which, in turn, depends on the incorporation of complex numbers into the analysis of group representations on vector spaces.

I have expressed this representation on a particular subspace of R^3, using the coordinates of R^3. One can also express this representation in matrices that correspond to the basis of V consisting of vectors v_1 and v_2. One uses the relationships $Tv_1 = v_2$ and $Tv_2 = v_1$ for T and relations $Rv_1 = \omega^2 v_1$ and $Rv_2 = \omega v_2$ for R. The matrices for T and R, corresponding to the basis consisting of v_1 and v_2, are:

$$T = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \text{ and } R = \begin{pmatrix} \omega^2 & 0 \\ 0 & \omega \end{pmatrix}$$

The representations for the remaining elements of S_2 can be obtained from these through matrix multiplication, as follows:

$$E = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad R = \begin{pmatrix} \omega^2 & 0 \\ 0 & \omega \end{pmatrix} \quad T = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

$$R^2 = \begin{pmatrix} \omega & 0 \\ 0 & \omega^2 \end{pmatrix} \quad RT = \begin{pmatrix} 0 & \omega^2 \\ \omega & 0 \end{pmatrix} \quad R^2T = \begin{pmatrix} 0 & \omega \\ \omega^2 & 0 \end{pmatrix}$$

Figure 27

I have now offered two irreducible representations of S_3. The first was the trivial action on R^1, in the form of the vectors in the onedimensional subspace U of R^3 consisting of vectors of the form (a, a, a). The second, I have just finished describing. Have I provided a complete list of irreducible representations of S_3?

No, there is one more, but it is a simple one. Let V be the one dimensional vector space $V = R^1$. Define a representation of S_3 on V by defining the actions of R and T as follows: For any vector v in V, let Rv = v and Tv = -v. In this action the group elements in the subgroup generated by R all act as the identity. The transpositions all act by multiplying by -1. In concrete terms, vectors in R^1 have only one coordinate. They all look like (x) where x is a number and (x) is the vector whose only coordinate is x. According to the formulas stated earlier, R and T act on (x) by R(x) = (x) and T(x) = (-x).

In this action, rotations are unimportant; only transpositions have any effect. If one remembers the example of the puzzle piece, one keeps track of which side is facing straight up; one cares about the flips. But rotational state doesn't matter. The trivial action of R corresponds to leaving, unchanged, the side facing straight up. The action of T, multiplying by -1, corresponds to a reversal. Transpositions multiply by -1; rotations multiply by 1. Appropriately if one applies T twice, multiplying twice by -1, the vector returns to its original value.

Earlier in the chapter, I introduced the idea of a quotient group, the group of transformations in S_3 considered without regard to rotation. It is this aspect of S_3 that is captured by this final representation of S_3.

This action arises, as well, from a slightly different represention, one more closely related to the quotient group. Consider the action on R^2 that I discussed a little earlier. Namely, suppose that R acts trivially on R^2, while T permutes the coordinates. In this representation, even T acts trivially on a subspace of vectors, specifically on vectors in the subspace U for which the two coordinates are equal. And, since both generators, R and T act trivially on U, the entire group, S_3, acts trivially on U.

But T also acts invariantly, yet non-trivially, on vectors in the subspace V of vectors with coordinates summing to zero. To see this concretely, notice that all vectors in V are multiples of the vector

$$\mathbf{v} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

To switch the coordinates of such a vector has the effect of multiplying that vector by -1. So the action of R and T on V is the same as the non-trivial action,

just discussed, on R^1. Like the previous example, transpositions multiply by -1; rotations multiply by 1.

We already saw that the non-trivial action of S_3 on R^1 is, essentially, the action of the quotient space in which one ignores the action of the rotations. It amounts to that action because only transpositions act non-trivially and because any two transpositions in S_3 have the same effect. In sum, the non-trivial action of S_3, on vector spaces of dimension 1, reflects, the structure of the quotient quotient of S_3.

There are, then, three irreducible representations of S_3. One is the trivial representation on a vector space of dimension 1 in which every element of S_3 acts as the identity. The second is the two-dimensional representation reflecting the full set of symmetries embodied in the group. And the third is the non-trivial onedimensional representation reflecting the symmetries of the quotient group.

This concludes my extended example of the representations of S_3. My presentation of group representations has emphasized three broad themes:

● Abstract finite groups are always realizable, both as permutation groups and as transformation groups acting on vector spaces.

● An important key to understanding the structure of a group is to study its representations.
● One classifies group representations by identifying their basic building blocks and identifying how group representations, in general, relate to these basic building blocks.

More broadly, a group is a system of measurements of symmetry. But identifications of symmetry are inherent in human conceptualization. Symmetry of some sort is involved every time one forms a concept. Whenever one can isolate a specific dimension or specific constellation of dimensions across which similar units differ, one has identified a kind of symmetry: One has identified a respect in which different things can be regarded as interchangeable and, yet, can also be related or compared. In essence, a transformation group is one way of measuring the differences between things that are similar from one perspective, yet different from another perspective. And, most fundamentally, that is what accounts for the ubiquity of transformation groups in mathematics.

Conclusion

Mathematics, as Ayn Rand put it, is the science of measurement. Measurement is the key to understanding mathematics. To understand mathematics as the science of measurement is to understand how mathematics relates to the world, to understand the precise sense in which mathematics is about the world.

We have seen aspects of this theme in every chapter. We learned that geometric abstractions provide an abstract focus on the *objects* of measurement. Geometry is an abstract way of looking at and studying *actual* objects and relationships in the world. It is *about* those relationships.

We recognized actual, real world triangles, circles, lines, etc., as the actual objects of geometric study. Euclid's postulates are all primitive measurements. And, as measurements, we must remember that their application to any concrete case is subject to the precision requirements of each case. Euclid's arguments are a form of indirect measurement, are recipes for a series of abstract measurements.

We analyzed magnitude, geometrically, as an object of numerical measurement. The essence of the Axiom of Archimedes consists in an implication, probably understood by Aristotle, that all magnitudes are measurable. We observed that relationships between numbers reflect quantitative relationships in the world among the quantities that they measure.

Following Euclid we learned just how the measurement of area, and the indirect measurement made possible by the laws of geometric proportion, both depend on the properties of parallel lines.

We discovered how irrational numbers relate to the world and why are they needed. We identified just what it means, in real world terms, to say that a Cauchy sequence converges to a real number and, in just what sense, the real number system is complete. And we sorted out the right and the wrong of the constructions by Dedekind and Cantor.

We found a reality-based account and rationale for the meaning and use of set theory in mathematics, emphasizing the importance of a proper hierarchy of mathematical abstraction. We reviewed, in stark contrast, the historical development, culminating in the conventional Zermelo-Fraenkel axioms of set theory. And we examined why, despite widespread acceptance of an incredible floating abstraction as a purported foundation, mathematics has survived.

We saw how the measurement perspective helps illuminate and integrate our understanding of vector spaces and linear algebra.

We studied a realm that is often thought to have little or no relationship to quantity, namely group theory. We saw how groups arise, why they are important, and, in just what sense the symmetry that they measure sits at the heart of the conceptual process itself.
Geometry is an abstract perspective on the objects of measurement. The real number system and transformation groups are systems of measurements. A mathematical argument is a series of abstract measurements, identifying quantitative relationships, pertaining, ultimately, to the world. Indirect measurement is purpose of mathematics and the source of its power.

[1] Moritz Epple, Chapter 10 "The End of the Science of Quantity: Foundations of Analysis, 1860 – 1910," In *A History of Analysis*, edited by Hans Niels Jahnke (Rhode Island, American Mathematical Society, 2003 hardback) p 291-324. Also see Jeremy Gray, *Plato's Ghost the Modernist Transformation of Mathematics*, 2008, Princeton, Princeton University Press, Chapter 3, "Mathematical Modernism Arrives", p 113-175, and Chapter 4, "Modernism Avowed," p 176-304

[2] Ayn Rand, *Introduction to Objectivist Epistemology Expanded Second Edition*, p 7-15 and also "Measurement, Unit, and Mathematics," p 184-203

[3] Mario Livio, *The Equation that Couldn't Be Solved*, New York, Simon and Schuster, 2005, Chapter 1, "Symmetry," p 1-28, and beyond. See also Hermann Weyl, *Symmetry*, Princeton, NJ, Princeton University Press, 1952 and Mark Ronan, *Symmetry and the Monster*, Oxford, Oxford University Press, 2006

[4] Livio, p 5-7

[5] Livio, p 4

[6] Ronan, *Symmetry*, "Galois: Death of a Genius," p 11-35. Also Livio, *Equation*, "The Romantic Mathematician," p 112-157 and Jorg Bewersdorff, *Galois Theory for Beginners*, Providence, RI, American Mathematical Society, for an elementary presentation of Galois's work

[7] Livio. Also John E. Maxfield, and Margaret W. Maxfield, *Abstract Algebra and Solution By Radicals*, New York, Dover Publications Inc.,1971, for an elementary introduction to group theory

[8] Rand, pp 11-15

[9] Maxfield, p 3

[10] Maxfield, p 3

[11] Maxfield, p 3

[12] Maxfield, p 3

[13] Maxfield, section on "Quotient Groups," p 70-77

[14] Robert D. Carmichael, *Introduction to the Theory of Groups of Finite Order*, New York, Dover Publications, Inc., 1937, p 3-26 regarding permutation groups

[15] Maxfield, p 3

[16] For a mathematician, something is a set if itcan be put into one-to-one correspondence with a set that exists within the scope of the ZF set theory axioms. For example, numbers are a set because one can identify them with a sequence of elements demonstrably contained in the ZF universe of sets. Leaving aside that any such identifaction steps outside the bounds of set theory, there is no such thing as the set of all

types of magnitudes, since the concept of magnitude is open-ended in a way that a set is not. Similarly, there is no such thing as the set of all finitedimensional vector spaces, unless one identifies any two vector spaces that are isomorphic to each other.

[17] Maxfield, p 58 on Cayley's Representation Theorem

[18] Ronan, especially Chapter 7, "Going Finite," p 79-87 and beyond. It is especially in regards to classifications that the ZF set theory axioms come into play. By specifying that any group has to qualify as a set per ZF, mathematicians have effectively specified a universe of discourse within which to consider the problem. (Note that this is only an issue for infinite groups.) A reality-based approach might, perhaps, have grounds to simply dismiss the classification problem, but otherwise needs a way to frame it. In the text I have given general reasons why classification is an important understaking.

[19] I. N. Herstein, *Topics in Algebra*, New York, Blaisdell Publishing Company, 1965, p 25, Regarding general structures, such as groups, Herstein says, "… always hoping that when these results are applied to a particular, concrete realization of the abstract system there will flow out facts and insights into the example at hand which would have been obscured from us by the mass of inessential information available to us in the particular, special case." He continues, "… the axioms which define them must have a certain naturality about them. They must come from the experience of looking at many examples; …". Having said all that, Herstein proceeds to provide a definition of a group, list some examples, and then plunges in with some lemmas. Herstein's standard text, aimed at serious mathematics majors, is excellent in many respects. To my current point, Herstein recognizes, officially, the importance of examples, but is, nonetheless, a representative example of the approach I'm talking about.

[20] Carmichael, Section 8, "Generators of Groups," p 30-31

[21] Richard H. Crowell and Ralph H. Fox, *Introduction to Knot Theory*, Boston, Ginn and Company, 1963, provide an advanced treatment on p 37-39

[22] Maxfield, p 189

[23] Lars V. Ahlfors, Complex Analysis, New York, McGraw-Hill Book Company, 1966, p 6-7

[24] Ahlfors, p 7

[25] Ahlfors, p 8

[26] Ahlfors, p 12-15

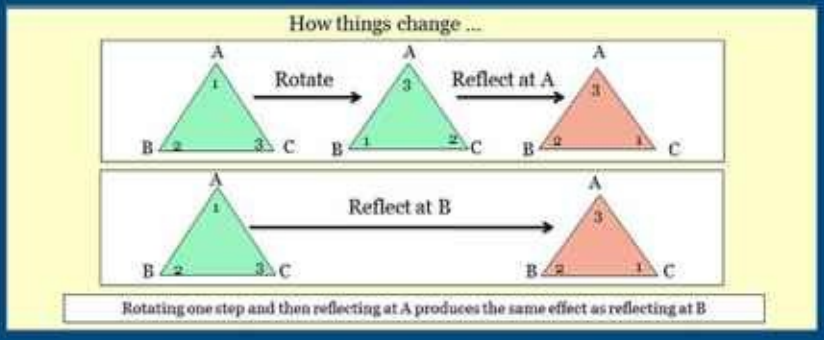
[27] Maxfield, p 58

[28] William Fulton and Joe Harris, *Representation Theory*, New York, SpringerVerlag, 1991, Chapter 1, p 3-11, an excellent text on a very advanced subject

[29] Fulton and Harris, p 6

What is mathematics about? Is there a mathematical universe glimpsed by a mathematical intuition? Or is mathematics an arbitrary game of symbols, with no inherent meaning, that somehow finds application to life on earth?

Robert Knapp holds that mathematics is about the world. His book develops and applies this viewpoint, first, to elementary geometry and the number system and, then, to more advanced topics, such as topology and group representations. His theme is that mathematics, however abstract, arises from and is shaped by requirements of indirect measurement. Eratosthenes, in 200 BC, demonstrated the power of indirect measurement when he estimated the circumference of the earth within about 16% without leaving Alexandria. Establishing geometric relationships, solving equations, finding approximations, and, generally, discovering quantitative relationships are tools of indirect measurement: They are the core of mathematics, the drivers of its development, and the heart of its power to enhance our lives.



Rotating one step and then reflecting at A produces the same effect as reflecting at B

Robert Knapp earned his Ph.D. in mathematics from Princeton University in 1973. He has published work on differential geometry and partial differential equations, and, after a year at the Insitute for Advanced Studies in Princeton, taught graduate and undergraduate mathematics at Purdue University. His study and appreciation of abstract mathematics and, simultaneously, his conviction that mathematics is about the world began in high school. Although he retired from the profession in the late 1970s, his study of the content, history and application of mathematics continues to this day. In recent years he has presented his unique perspectives on geometry and the number system in a series of lectures at Objectivist Summer Conferences organized by the Ayn Rand Institute.