AN EXPLORATION OF CONNECTIONS BETWEEN HIGH SCHOOL ALGEBRA AND ABSTRACT ALGEBRA

by

CHEN TIAN

(Under the Direction of Denise S. Mewborn)

ABSTRACT

This thesis illuminates some connections between high school algebra and abstract algebra. The concepts of function (e.g., homomorphism and 1-1 correspondence) and algebraic structure (e.g., ring, field, and group) are compatible with each other. This thesis mainly deals with (1) functions from an overall structural perspective, (2) the solvability of some polynomial equations in certain algebraic systems, and (3) the real numbers from both constructive and axiomatic approaches. Throughout this thesis I provide possible questions or tasks that teachers can use to challenge students to think of mathematics in a bigger picture as well as help students develop an admiration for some aesthetic dimensions of mathematics. I identify some questions that high school students might ask and provide answers are given from a relatively advanced mathematical standpoint to illuminate the connections between high school algebra and abstract algebra. The notion of limit in calculus and the interaction between algebra and geometry are also rendered in several contexts.

INDEX WORDS: high school algebra, abstract algebra, connections, algebraic structure, isomorphism, the real numbers, polynomials

AN EXPLORATION OF CONNECTIONS BETWEEN HIGH SCHOOL ALGEBRA AND ABSTRACT ALGEBRA

by

CHEN TIAN

BS, Xi'an University of Architecture & Technology, P. R. China, 2007

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF ARTS

ATHENS, GEORGIA

2010

AN EXPLORATION OF CONNECTIONS BETWEEN HIGH SCHOOL ALGEBRA AND ABSTRACT ALGEBRA

by

CHEN TIAN

Major Professor: Denise S. Mewborn

Committee: Roy Smith
Jeremy Kilpatrick

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2010

DEDICATION

To

my dearest parents who raise, enlighten and support me without reservation but with boundless love and wisdom and are a great influence on making me the person I am today;

my teachers and friends whom I have been fortunate to know, to learn from, and to be supported by;

my bittersweet journey at the University of Georgia (UGA), through which I have been growing up a lot and gaining a different perspective on life;

myself whom I am getting to know better, whose faults I should be responsible for, and whose success I am proud of;

and

my bright future ...

ACKNOWLEDGEMENTS

First and foremost, my heartfelt gratitude goes to my thesis committee members. Dr. Denise Mewborn is my committee chair and also my academic advisor. I truly appreciate her guidance and support in each step of developing my thesis. I get the impression that her advising is pithy and encouraging but not pushy. She is the one who started my unforgettable connection with UGA Math Education, threaded through and gave a hand to most of the decisive moments in my life at UGA, and helped me settle down with the current topic of my thesis. Given that English is not my native language, she even spent time in editing my thesis. It is so nice of her. It is lucky to have such a pretty, elegant, and sage major advisor like her once in life. Dr. Roy Smith is a "maverick" mathematician and mathematics educator. I have great respect for him. I have gained a solid understanding in some topics in mathematics from him and have been heartened by his enthusiasm for teaching inspirational mathematics classes. The classes I took from him and the communication I had with him played a highly contributory role in the writing of this thesis especially the mathematical ingredients. Knowledgeable Dr. Jeremy Kilpatrick's questioning during my committee meeting was instrumental in promoting me to reorganize my thesis into the current shape. His other detailed and insightful advice on my thesis was also enlightening. The reading of his doctoral seminar course (which I audited several times) also encouraged me to reflect upon my own philosophical concerns about mathematics education in my thesis. My oral defense turned

out to be like a “chat” through which I continued to learn more from the professors. When the time was over, I wished it had lasted longer.

I now express my special thanks to those who were so generous with answering my questions or giving different help for my thesis writing: my super supportive Mom, Dr. Clint McCrory and his Abstract Algebra II course (MATH 6010), Dr. Edward Azoff, Dr. Sybilla Beckmann, Dr. Leslie Steffe and his EMAT 7080 course, the textbook (1996) written by Dr. Theodore Shifrin, EMAT 7050 course offered by Dr. Anna Conner, EPSY6800 instructed by Dr. Marty Carr, Abstract Algebra I (MATH 6000) and Foundation of Geometry I (MATH 7200) courses both taught by Dr. Roy Smith, and the flower of other EMAT courses I have taken, Kelly Edenfield (who is the first friend I met at UGA and to whom most of my academic questions went first), Dana TeCroney, Laura Nunley, Zandra de Araujo, Anne Marie Marshall (who is my “elder” good friend and whom I sometimes call “Anne Mama”).

I also want to say thank you to my other buddies: Natalie Lord, John Lubeski, Kyla Gaffney, Larousse Charlot, Megan Dickerson, Jonathan Beal, Wenyuan Xiao, Wenfei Liu, Chan Zhou, Rui Kang, Fen Gui, Xiaohua Wei, Yubing Wan, Xiang Gao, Rongrong Shi, ... I hate that it is impossible to list all my friends here. We had such a memorable time together. You were the sunshine of my life at UGA, and I know the warm light will continue shining across the borders of countries.

Last but not least, I am grateful for life itself. Life is not fair but is still fabulous and fascinating. Life is “isomorphic” to an algebra curriculum, and I am here to learn and explore. After interacting with all the lessons, solving most of the problems, and passing

all my tests, expecting uncertainties, I eventually realize that what I have been learning is just basic algebra.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

1. Introduction

1.1 Background and motivation

According to U. S. President John F. Kennedy (1960), a special study by the Joint Atomic Energy Committee of Congress pointed up the responsibility of U.S. school mathematical and scientific education:

> The teaching of the physical sciences and mathematics in our secondary schools has declined; about half of those with talents in these fields who graduate from high school are either unable or uninterested in going to college; and of the half who enter college, scarcely 40 percent graduate. The task of reversing these disturbing trends is in large measure up to our public schools and their teachers. It is up to our teachers' colleges and their graduates. (para. 5)

Middleton and Spanias (1999) pointed out that

> National assessment data from the 1980s (Carpenter, Corbitt, Kepner, Lindquist, & Reys, 1981; Dossey, Mullis, Lindquist, & Chambers, 1988) have indicated that American children tend to enjoy mathematics in the primary grades but that this level of enjoyment tends to fall dramatically when children progress into and through high school. In addition, although students feel that mathematics is important, the number of students who want to take more mathematics in school is declining steadily (Dossey et al., 1988). These statistics seem alarming when coupled with the fact that children do not possess the mathematical knowledge that they will need to function smoothly in our increasingly technological society. (p. 65)

"In the 1960s, we sought to motivate the mathematics; in the 1990s, we seek to motivate the students" (Foley, 1998, p.87).

Encouragingly, "[e]very child is now promised a college education. … A new surge of students is entering higher education expecting faculty to prepare them in their discipline and provide the background for future careers." (MAA, 2001, p. 1). Venezia, Kirst, and Antonio (2003) cited the national statistic that 88% of 8th graders expect to

participate in some form of postsecondary education, and approximately 70% of high school graduates actually do go to college within two years of graduating. They also reported that 88% of all students (cutting across racial and ethnic lines) surveyed for the Stanford University's Bridge Project intend to attend some form of postsecondary education.

Unfortunately, most high schools have not met the students' heightened educational aspirations and are not doing a good job in preparing students well for college and the labor market. The content of the last 20 years of mathematics has been ambiguous, and as a consequence, a sizable percentage of students graduated from the so called "shopping mall high school" (Chazan, 2008, p. 20) with nonacademic training in mathematics. In addition, Venezia, Kirst, and Antonio (2003) reported that many students do not have a good sense of what is expected of them in college, and most educators do not know how to help students gain an understanding of those standards. The courses and tests students are taking to graduate from high school and attend college have little to do with the academic expectations that students face in their first year in college. A study conducted for Achieve, Inc. (2005) showed that substantial proportions of high school graduates identify gaps between the education they received in high school and the overall skills, abilities, and work habits that are expected of them today in college and in the workforce. Additionally, 42% college students feel that there are some gaps in their mathematics preparation (13% large gaps), and 300 interviewed college instructors (with margin of error ±5.6%) and 400 interviewed employers (with margin of error ±4.9%) estimate, respectively, that 42% and 45% of recent high school graduates are not adequately prepared for the skills and abilities they need to advance beyond entry level jobs. The

survey also provides data evidence that, in contrast to the opinion that high school students would resist changes that would force them to work harder, those current high school graduates, knowing what they know now, say they would have worked harder. Like their college instructors and employers, those high school graduates say higher expectations, more rigorous curricula and tougher requirements for high school graduation would leave them better prepared for the real challenges they are now facing in college and the work world.

The NCTM *Curriculum and Evaluation Standards* (1989) had "an overarching theme of 'Mathematics as Reasoning'" (Foley, 1998, p. 87) as evidence by the expectation that "high-school students should learn about … mathematical systems and their structural characteristics, … difference equations, the complex number system, elementary theorems of groups and fields, and the nature and purpose of axiomatic systems" (p. 87). Also according to Foley (1998), the American Mathematical Association of Two-Year Colleges' *Crossroads in Mathematics* (1995) contains similar calls for the content of college-preparatory mathematics.

From the angle of the education of mathematics teachers, Ferrini-Mundy and Findell (2001) expressed the same concern:

> For secondary mathematics teachers, it is ironic that, except for occasional concepts that might be called upon in calculus, the entire four years of an undergraduate mathematics major address content that is, on the surface, unrelated to the topics of the high school curriculum. … More substantially, the kinds of integration of mathematical ideas and connections that are necessary in teaching a coherent secondary program are unlikely to be obvious to students on the basis of their undergraduate program. (p. 33)

The issues identified above call for effective instruction in mathematics classrooms to help students see connections in mathematics and help them build bridges as they

make different transitions in mathematics learning. In this thesis I will concentrate on this idea by introducing some connections between high school algebra and abstract algebra to help high school teachers and students see the connections between them. I hope to illustrate challenging but doable tasks to engage high school students in algebra and mathematical thinking. To be clear, by saying "a task" here in this thesis, I mean an activity or a problem on which students may be asked to work or the directly related raw material from which teachers could construct a specific task for students. For instance, to find the quantitative pattern between the sides of a regular polygon and the rotations of symmetry (c.f., Chapter 4.2) could be a task, or to give a reason why a polynomial can be factored or not in a certain context could also be one (c.f., Chapter 3 and 5.3). Some of the questions listed in next section (Chapter 1.2) could also be developed into classroom activities, such as to develop a way to enumerate the rational numbers (c.f., Question 21), to find how many ways are there to label an irregular triangle using the different letters of the alphabet (c.f., Question 12), to use Fermat's Little Theorem to create your own coding method for integers (c.f., Question 11), to check whether a big number is a perfect square (c.f., Question 4), to show whether $0.999\ldots = 1$ (c.f., Question 14), and to find out a few differences between the rational numbers and the real numbers (c.f., Question 25).

Henningsen and Stein (1997) classify the cognitive demands of mathematical tasks into: (i) memorization; (ii) the use of formulas, algorithms, or procedures without connection to concepts, understanding, or meaning; (iii) the use of formulas, algorithms, or procedures with connection to concepts, understanding, or meaning; (iv) and cognitive

activity that can be characterized as “doing mathematics”, including complex mathematical thinking and reasoning activities such as making and testing conjectures, framing problems, and looking for patterns. The last two ((iii) and (iv)) are considered as placing high-level demands on students, and the tasks (including the relevant raw material) I explore in this thesis fall into these categories.

High-level thinking involves being able to connect multiple ideas and stretch ideas beyond the immediate context, which is challenging to students. Students who think mathematically with the help of a teacher or peers may be able to confront a problem where the solution path is not immediately obvious and figure out what to do (perhaps in more advanced way). They are also able to use reasoning to decide if their conjecture or solution is sensible or not. The tasks I present in this thesis are meant to foster students’ disposition of curiosity and perseverance toward (hopefully, deep and significant) mathematics.

Another of my motivations for pursing this thesis topic is the general opinion of students enrolled in abstract algebra courses that there is little or no connection between abstract algebra and high school algebra. For example, a student posed the following question about abstract algebra on Yahoo! Answers (2009) “…If I did average on algebra in my high school, how well will I be able to understand this?”. The main answers to that question on line were: (1) “There's no comparison between high school algebra and abstract algebra. The former is mostly just computation and graphs. The latter is algebraic structures and proofs about them.” (2) “You'll probably want to be fairly well grounded

in discrete math, set theory, and number theory before taking abstract algebra on. It's almost completely unrelated to high school algebra, so go into it expecting that kind of math." (3) "[T]his killed my university degree in math, but looking back, the error I made was to make it "math" (like calculus; this and this equals this). It's more logic than mathematics."

Abstract algebra can be "a course of an encyclopedic nature dealing critically with the field of elementary mathematics from the higher standpoint" (International Commission on the Teaching of Mathematics, 1911, pp. 13–14, as cited in Ferrini-Mundy & Findell 2001, p. 32). In mathematics education, abstract algebra courses can help pre-service and in-service teachers refine and expand middle and high school algebra concepts and can provide experiences in posing questions that encourage student-directed learning in the exploration of the mathematics curriculum. Some textbooks (e.g., Shifrin, 1996) have been written with this purpose in mind. Algebra courses should motivate students to ask questions such as "What algebraic setting is given?", "If I change the parameters or initial conditions, how will that affect the problem and its framework and solution?", or "How do these algebraic and geometric ideas mesh?". I believe drawing connections from abstract algebra to high school algebra can help enhance students' algebraic understanding or even mathematical thinking as mathematics learners and "help them become better mathematical doers and thinkers" (Henningsen & Stein, 1997, p. 524).

The National Council of Teachers of Mathematics [NCTM] (2000) encourages making connections, noting that

> Instructional programs from prekindergarten through grade 12 should enable all students to—
> • Recognize and use connections among mathematical ideas
> • Understand how mathematical ideas interconnect and build on one another to produce a coherent whole
> • Recognize and apply mathematics in contexts outside of mathematics (p. 354)

Some teachers do realize that they have "come to believe that the act of building connections and relationships is at the heart of mathematical proficiency" (Boaler & Humphreys, 2005, p. 11). Boaler and Humphreys (2005) commented that

> '[i]f curriculum and instruction focus on mathematics as a discipline of connected ideas, students learn to expect mathematical ideas to be related' (NCTM, 2000, p. 275). As students make these connections and develop understanding of these relationships for themselves, the fabric of their mathematical proficiency becomes ever more flexible and sturdy. (p. 11)

"Teaching mathematical topics as disconnected entities, or as a sequence of 'tricks of the day', may lead to high quiz scores at the end of the week but rarely will lead to long-term understanding (Steen 1999). Rather, in-depth explorations of the relationships among representations and ideas help students develop a more reliable and sustainable capacity to use, transfer, and understand mathematical ideas and procedures" (Martin, 2007, p. 26). Thus, effective mathematics teachers actively engage students in tasks that enable them to see mathematics as a coherent and connected endeavor rather than as a series of disconnected rules and procedures that they must memorize (Martin, 2007). A further perspective expressed by Lang (1985a) was that

> In most school books, the topics are usually treated in a way which I find incoherent. They pile up one little thing on another, with out showing the great lines of thought in which technique can be inserted, so that it becomes both appealing and meaningful. They don't show the great mathematical lines, similar to musical lines in a great piece of music. And it's a great pity, because to do mathematics is a lively and beautiful activity. (p. xi)

Again, my goal in writing this thesis is to illuminate some connections between abstract algebra and high school algebra, and the way I organize my thesis also respects that mathematics should be viewed from a connected and exploratory standpoint. All the abstract algebra books I have seen (e.g., Shifrin, 1996; Birkhoff & MacLane, 1941) are written in a quite mathematically sophisticated style tending to present concepts and proofs in an articulate, but somewhat succinct, rigorous, elegant way, sometime even leaving parts of some arguments for the student to fill in. To make the illumination as plain as I can, in this thesis I first elaborate an idea from abstract algebra and then identify related mathematical ideas with which high school students may be able to grapple. In some cases I start with a context with which high school students are familiar and then talk about the abstract algebra ideas entailed in it. These connections and the tasks embedded in them involve analyzing structures, dealing with functions, making choices about representation, and manipulating expressions, all of which are intrinsic to mathematics, and particularly to algebra (as a special language itself having its own set of grammatical rules that are not intuitive but must be learned and practiced (MacGregor & Price, 1999)). Although this thesis mainly involves algebra knowledge, sometimes I also use the ideas from calculus and geometry, because I consider calculus as "algebra with limits" and geometry as "algebra with figures." I give the proofs for most theorems in a relatively plain and narrative way, which is expected to be easier for high school students to follow than reading those succinct ones written by high-achieving mathematicians.

1.2 Can we answer these questions for our students?

The following list of questions that high school students might ask are answered in this thesis, and this thesis also provides a direction (with some possible questions or tasks as well as an awareness of some aesthetic aspects of mathematics) for teachers to challenge the students to think of mathematics in a bigger picture.

Questions students might ask include:

(1) What is a function? (Throughout this thesis, especially Ch.2 (def. in 2.1))

(2) Where are exponential and logarithmic functions used? (2.2)

(3) Why do complex roots of a polynomial with real coefficients come in conjugate pairs? (2.3)

(4) How can I check whether a very big number is a perfect square? (2.4)

(5) How do you know $34x^4+16x^3-8x-103=0$ has no integer solution? (2.4)?

(6) You said a yo-yo and the train of a peacock could be mathematical metaphors. What are they? (2.4)

(7) I keep forgetting the binomial formula. Is there a way to help me memorize it or develop it by myself? (3)

(8) What does it mean to solve an equation? (2.1; 3; 5.3)

(9) How do you know these equations $8x^3-27x^2-6=0$, $x^3-32=0$, $x^4+x^3+x^2+x+1=0$ have no rational solutions for x? (3; 5.3)

(10) You said if we can factor a polynomial $f(x)$ [into $Q[x]$] then we can also factor $f(x+1)$ [into $Q[x]$]. Why? (3)

(11) I always wonder how mathematics is useful in our life. I heard that Fermat's Little Theorem is the basis of modern cryptography. And I remember when we learned the

binomial theorem, you said we can use it and mathematical induction and the meaning of divisibility to prove Fermat's Little Theorem. How? And how is it related to coding? (3)

(12) How many ways are there to label a triangle using the different letters of the alphabet? (4.1)

(13) If I rotate a square, how many different ways are there to take it back to its original position? What about a pentagon or a cube? (4.2; 4.3)

(14) Does 0.999… really equal 1? Then why does my calculator show $(0.999999999)^2 = 0.999999999$? (3; 5.2; 6.4)

(15) Can you prove there are an infinite number of natural numbers? (5.1)

(16) Why is the number of even or odd numbers not half of the natural numbers? (5.1)

(17) Are there infinitely many prime numbers? (5.1) How is the concept of prime numbers useful? (3)

(18) I believe that every positive integer greater than one can be factored into positive prime factors. But is there only one unique way to factor? (5.1)

(19) What is a rational number? (5.2; 5.4) Particularly, what is $1/n$? (5.2; 1.3)

(20) Are there as many rational numbers as natural numbers? (5.1)

(21) Now I see why there are as many rational numbers as natural numbers, because we can enumerate the rational numbers by finding a way to list them each one corresponding to a different natural number. Can we enumerate the real numbers like that? (5.1; 5.3)

(22) How do we know there is a one-to-one correspondence between the real numbers and the points on a straight line? (5.2)

(23) What is the previous (or next) number before (or after) 1/2 [in a specific number system]? (5.2)

(24) How do we add (or multiply) two (positive) infinite decimals? (5.2)

(25) What differentiates the real numbers from rational numbers? Is it that a rational number can be written in terms of a pair of integers whereas an irrational number cannot? (5.2; 5.3; 6.3)

(26) Is the "*x*" in a polynomial a "variable" or an "unknown"? (5.3)

(27) Why do you say that for this equation $\frac{x^2-25}{5-x}=-(x+5)$ we may or may not need the "as long as x does not equal 5" part? (5.3)

(28) What is an algebraic number? Are algebraic numbers enumerable? (5.3)

(29) What is meant by $\sqrt{2}$? Is it a real number? How would you describe it? How do you know $\sqrt{2}$ is not equal to 2? (5.2; 5.3)

(30) Is $\sqrt{2}$ a rational number? / Is there any rational number that satisfies $\sqrt{2}$? (5.3; 6.3)

(31) If there is not a rational number that satisfies $\sqrt{2}$, is there a number in some other number system that will satisfy $\sqrt{2}$? (5.2; 6.4)

(32) Are algebra and geometry related to each other? (mainly 4; 5.4)

(33) I know how to construct a segment of length $\sqrt{2}$, but then can I use a straightedge and compass to construct a segment of length $\sqrt[3]{2}$? (6.4)

(34) We are told $1\neq 2$ [where 2=1+1 as natural numbers]. Is it really true? Why is that important? (6.4)

(35) Why can we not divide a number by 0? What about 0/0? (6.4)

(36) Why does a negative times a negative equal a positive? (5.2; 6.4)

1.3 Algebraic structure

Rickart (1996) addressed two aspects of teaching and learning school algebra: "the degree of understanding of the background subject matter leading up to the immediate subject of interest, and … the formalism associated with the subject" (p. 293). However, sometimes the teaching is reduced to nothing more than teaching the formalism; mastering some formal algebraic operations without association with the concept of an algebraic system somehow replaces the desired algebra concept with the relatively superficial structure of the associated formal language. When a student's understanding or conception of the underlying algebraic structure is inadequate it will be problematic for him/her to deal with the more advanced topics in mathematics (Rickart, 1996). "Teachers of mathematics in Grades 11-14 must understand algebraic groups, rings, fields, and the associated theory. … in keeping with the NCTM curriculum standards for college-intending students, high-school teachers need to be able to convey such understanding to their upper level students" (Foley, 1998, p. 88). "The teacher's objective in algebra should include a constant awareness of the desirability of helping the student to supply or develop the basic understanding of an algebra system" (Rickart, 1996, p. 294). Therefore, in this section, I elaborate on the big picture in terms of some basic algebraic structure concepts (e.g., ring, field, group) in abstract algebra. These concepts appear throughout this thesis and are compatible or bound up with the concepts of functions (e.g., isomorphism). In particular, ring and field theory deals with solutions to polynomial equations in a precise and systematic way.

Therefore, at first, I consider the following nine basic algebraic laws (for *every* element in the given set):

(1) Additive Commutative Law

(2) Additive Associative Law

(3) Multiplicative Associative Law

(4) Additive Identity Existence

(5) Multiplicative Identity Existence

(6) Multiplicative Commutative Law

(7) Distributive law

(8) Additive Inverse Existence

(9) Multiplicative Inverse Existence

Integers (denoted by Z) can be added and multiplied, subject to the first eight algebraic laws. All high school students know the fact that an integer plus its inverse is 0. 0 here is the additive identity of Z, satisfying Law (4) above. And we can find another integer for an integer such that the sum of them equals zero, this shows Law (8) for Z. The product of 1 and any integer is still the integer itself. This means 1 is the multiplicative identity of Z, satisfying Law (5) above. But nobody can find another integer for an integer (except 1 and -1) such that the product of them equals 1. So integers do not have Law (9).

However, we all know rational numbers Q, in terms of $\frac{m}{n}$ with $m, n \in N$ (natural numbers), have not only the first eight algebraic laws, but also have the ninth law. So now we see there is some difference between Z and Q with the operations + and ×.

Here we define a **structure** as a set of elements with respect to certain operation(s) satisfying some rules or axioms. We call a set of numbers or even objects with operations + and × a **ring** if they have the laws (1)～(5), (7), and (8), and call a ring “commutative

ring" if it also satisfies Law (6), and call a commutative ring a **field** if Law (9) also works for every nonzero element in the set of numbers or objects.

We can see the specialty of the 9th law, so we call a nonzero element of a ring that has a multiplicative **inverse** a **unit**. Hence we can define a ring is a field if every nonzero element of the ring is a unit.

If there are nonzero elements a and b of a ring such that $ab = 0$, then we say x and y are **zero-divisor**s in the ring. We define that a nontrivial ring (i.e., $0 \neq 1$) is a **domain** if it contains no zero-devisors, and that a commutative domain is an **integral domain**. Note that the trivial ring is commutative. And note that every field is an integral domain, because if a is a unit and $ab = 0$, then $b = 1b = (a^{-1}a)b = a^{-1}(ab) = a^{-1}0 = 0$; not every integral domain is a field, for example, the integers Z.

In mathematics history, some great mathematicians (e.g., Galois, Lagrange, Cauchy, and Abel) worked on the study of solutions of polynomials. They found only considering some laws of a set of objects for one operation was very helpful and interesting, so the concept of group arose.

A **group** is a set G with an operation $\cdot$ (note that we can define the operation as we need), such that

(0) for all $a, b \in G$, $a \cdot b \in G$ (Closure)

(1) For all $a, b, c \in G$, $(a \cdot b) \cdot c = a \cdot (b \cdot c)$ (Associativity)

(2) There is an element $e \in G$, such that for all $a \in G$ $e \cdot a = a$ and $a \cdot e = a$ (Identity)

(3) For all $a \in G$ there exists $a^{-1} \in G$ such that $a \cdot a^{-1} = e$ and $a^{-1} \cdot a = e$ (Inverse)

If a group G satisfies the law of commutativity, i.e., for all $a, b \in G$, $a \cdot b = b \cdot a$, we say the group G is abelian (or commutative). For the interest of some school students, the

adjective “abelian” was derived from noted Norwegian mathematician Niels Abel (1802-1829) who invented (independently of Galois) an extremely important invaluable branch of mathematics known as group theory.

Thus, we know $(Z,+,\times)$ is a ring, but not a field. But now $(Z,+)$ is a group if we only consider one operation +, while $(Z,\times)$ is not (but do not forget that the set of units in a commutative ring forms a group under multiplication, for instance, $(\{1,-1\},\times)$). One more example: $(Q,+,\times)$, or $(R,+,\times)$, or $(Z,+,\times)$ is a ring, and also is a field. For instance, we all know that $0+a=a$, $1\times a=a$, $a+(-a)=a$, and $a\cdot\frac{1}{a}=1$ for all $a\in R$. If we use the terminologies in abstract algebra, like we did for Z, 0 is the additive identity, 1 is the multiplicative identity, $-a$ is the additive inverse of a, and $\frac{1}{a}$ is the multiplicative inverse of a, for ring $(R,+,\times)$. Similarly, $(Q,+)$, $(Q,\times)$, $(R,+)$, $(R,\times)$ are all groups. For instance, when students say the product of two rational numbers is still a rational number, they essentially use the closure of group $(Q,\times)$.

However, things will become less regular as we move through various number systems. Moreover, the transition from basic number systems to more sophisticated algebraic systems is not a minor step in students’ understanding. Thus, students need to study properties and possibilities of important algebraic structures associated with functions in order to discern their similarities and differences, such as isomorphism (homomorphism plus bijection and some other relevant concepts such as congruence and symmetry in certain contexts), modular operations (or quotient ring), ordered field, field extension, typical polynomial rings and fields, and constructively important properties for the real numbers.

2. Isomorphism

Most of the current research in mathematics education treats functions as "a certain type of procedure" or "a process that transforms input values to an output value in a predictable way" (Cuoco, 1990, p. 19). This conception puts the accent on the specific foreseeable relation between individual elements of two sets rather than a powerful relation connecting two structures, the perspective from which the notion of isomorphism is construed in this chapter. I develop the discussion in four sections with several different contexts.

The first section mainly defines an isomorphism by its two properties associated with two algebraic structures: it preserves algebraic operations and it corresponds elements one-to-one and onto. A few examples are included to illustrate the concepts.

Then, in the second section, I connect the notion of isomorphism directly to logarithmic and exponential functions and some of the applications, the important concepts in high school algebra which are virtually embedded in a larger picture of mathematics.

In the third section I introduce the concepts of field extension and automorphism (a special isomorphism), as well as symmetry, so as to give an in-depth discussion about the proof of the conjugate zeros theorem from an advanced standpoint.

The fourth section centers around modular arithmetic, because as an equivalence relation it is a good example of homomorphism. In particular, the concept of

homomorphism is a weaker variation of isomorphism but with even more powerful applications than isomorphism because there are more instances of homomorphisms than isomorphisms, and a homomorphism sometimes simplifies the problem whereas an isomorphism does not. At the end of this section, in order to inspire some students' interest in further pursuit of mathematics I briefly introduce the Fundamental Group Homomorphism Theorem and its sophisticated connection to the Cartesian Product.

In addition, the topics in the other chapters are also related to isomorphism in some way. Because they have different points of focus, however, I organized them into individual chapters.

2.1 What is an isomorphism?

I had never seen the word "isomorphism" when I was in school. That word must look complicated and maybe a little bit scary to almost every school student, and even to me when I, as an international student, was starting to take abstract algebra course. If we search the word online, we will see that in Greek "isos" means "equal" and "morphe" means "shape." But what does "equal shape" mean, mathematically? High school students should be familiar with the term "isosceles triangle," which shares the prefix "iso" with "isomorphism". The following is a description of the purpose of the study of isomorphism:

> Isomorphisms are studied in mathematics in order to extend insights from one phenomenon to others: if two objects are isomorphic, then any property which is preserved by an isomorphism and which is true of one of the objects is also true of the other. If an isomorphism can be found from a relatively unknown part of mathematics into some well studied division of mathematics, where many theorems are already proved, and many methods are already available to find answers, then the function can be used to map whole problems out of unfamiliar territory over to "solid

ground" where the problem is easier to understand and work with. ("Isomorphism", n.d., "Purpose")

Students may get some flavor of the magic role of isomorphism in mathematics but must still feel uncomfortable with the concept of isomorphism. I could give a few examples with which high school students are familiar as motivation and then introduce the relevant technical definitions used in abstract algebra. But after giving a second thought of this I decide to state those definitions at the beginning, because the notion of isomorphism itself is pretty fundamental in algebra (even in mathematics), and it is worth an independent consideration like we do for algebraic structures while both of them are compatible with each other. I think after correctly understanding the basic definitions it will be easier for students to understand the upcoming examples better.

There are two basic and important isomorphisms in abstract algebra – ring isomorphism and group isomorphism. Before we define isomorphism we need to first define function, and then homomorphism and bijection as special cases of functions.

A **function** $\phi : A \to B$ is a relation or mapping between A a given set of elements called the **domain** and B a set of elements called the **codomain**. The function associates each element (often denoted by the letter x and called the **independent variable** or **argument** or **input** of the function) in the domain with *exactly one* element (often denoted by the letter y and called **dependent variable** or **value at** x or **image of** x or **output** of the function) in the codomain. The elements so related can be any kind of things, for example, numbers, polynomials, functions themselves, sets, real-life objects, words, etc., but typically mathematical objects. We can think of a function as a set of ordered pairs (x, y), but usually we use equation $y = \phi(x)$ to define a function.

Incidentally, we define an **equation** as a mathematical statement that asserts the equality of two expressions. It usually involves some unknown number(s), and solving it means finding the number(s) that make(s) the statement true. But sometimes depending on the context we may or may not be able to find the solution(s), several instances of which will be shown in the following chapters (especially Chapter 2.3, 2.4, 3 and 5). For any function $\phi : A \to B$, the **kernel** of the function is defined by $\ker \phi = \{x \in A : \phi(x) = e'\}$ where e' is the additive identity of B (say, 0 in R) or multiplicative identity of B (say, 1 in the R), i.e., the kernel gives the elements from the original set A which are mapped to identity in set B by the function. So $\ker \phi$ is a subset of A. The related **image** of this function is defined by $\mathrm{Im}\, \phi = \{\phi(x) : x \in A\}$, i.e., the image is the elements in set B, which are got by mapping the elements originally from set A by the function. So $\mathrm{Im}\, \phi$ is a subset of B. Usually the image of the domain A of the function ϕ is also called the **range** of the function. The function is a **surjection** if the range of the function is equal to the codomain B. The function is an **injection** if it maps distinct arguments x to distinct images of x under ϕ. These concepts should not be perplexing to high school students, but they may look pretty technical to most students. Teachers can make efforts to engage students in progressive practice with various examples of functions.

A **ring homomorphism** is a function between two rings, which preserves the operations of addition and multiplication.

More precisely, if (R, +, ·) and (S, +,·) are rings, then a ring homomorphism is a function $\phi : R \to S$ such that for all $a, b \in R$,

（1） $\phi(a + b) = \phi(a) + \phi(b)$

（2） $\phi(a \cdot b) = \phi(a) \cdot \phi(b)$

(Note that the so-called “addition” and “multiplication” on the left-hand side are in R, while those on the right-hand side are in S. They can be any operation in the groups as defined.)

A **group homomorphism** is an operation-preserving function between two groups. More precisely, if (G, +) and (K, ·) are groups, then a group homomorphism is a function $\phi : G \to K$ such that for all $a, b \in G$,

$\phi(a + b) = \phi(a) \cdot \phi(b)$

(Note that the so-called “addition” operation on the left-hand side is in G, while the so-called “multiplication” operation on the right-hand side is in K. They can be any operations in the groups as defined.)

$\phi : A \to B$ is a **bijection** (namely, one-to-one (injective) and onto (surjective) correspondence) if and only if for every y in B there is a unique x in A with $\phi(x) = y$.

Thus we say an **isomorphism** is a **bijective homomorphism**.

Before we move to the examples that students must have been anxious to see, please allow me to introduce two useful lemmas.

Lemma 2.1 (c.f., Shifrin, 1996, p. 125):

A ring homomorphism $\phi : A \to B$ is one-to-one (not necessarily onto) $\Leftrightarrow$ $\ker \phi = \langle e \rangle$

where $\langle e \rangle$ is the **ideal generated by** (defined in Shifrin, 1996, p. 117) the additive identity e of A.

Lemma 2.2 (c.f., Shifrin, 1996, p. 182):

Given $\phi: A \to B$ a group homomorphism, $\ker\phi$ is a subgroup of G, and ϕ is one-to-one (not necessarily onto) $\Leftrightarrow$ $\ker\phi = \{e\}$, where e is the identity of A.

(To verify these two lemmas will be a good exercise for students to get more familiar with the concepts of subgroup, identity, inverse, homomorphism, injection, and kernel, and also the interaction between the necessary condition and the sufficient condition in the proof.)

Now let us check whether the typical quadratic function $\phi: R \to R$ defined by $y = \phi(x) = x^2$ is an isomorphism as follows.

$(x_1 + x_2)^2 \neq {x_1}^2 + {x_2}^2$; $(x_1 \cdot x_2)^2 = {x_1}^2 \cdot {x_2}^2$, so ϕ is only a group homomorphism from group $(R, \cdot)$ to $(R, \cdot)$. But $\ker\phi = \{1, -1\} \neq 1$, since $(\pm 1)^2 = 1$. So by **Lemma 2.2** above ϕ is not injective. Hence ϕ is not an isomorphism.

Remember that if there exists an isomorphism between two rings R_1 and R_2, then we say R_1 and R_2 are isomorphic. Denote $R_1 \cong R_2$. So to speak, the two rings essentially are "the same", similar for groups. More generally, isomorphic structures, despite of notations, essentially are identical. In other words, two structures having "different outfits" can be essentially "the same" if they are mathematically isomorphic.

Does the above example mean that the group R is not isomorphic to itself? Definitely not! But what was wrong there? Please be careful that we said "if there exists an isomorphism". In other word, if we could find an isomorphism which maps R to R then we proved that R is isomorphic to itself. One obvious ring/group isomorphism is

$\phi : R \to R$ defined by $\phi(x) = x$. Incidentally, this is essentially consistent with rigid motions in Cartesian plane over R.

To help students get acquainted with this basic technique we consider the following examples. By the way, I suggest that teachers assign students to collect two or three functions to decide if each of them is a homomorphism or even an isomorphism and give the reasons.

- $\phi : Z \to Z$

(1) If $\phi(x) = 5x$

$\phi(x_1 + x_2) = 5(x_1 + x_2) = 5x_1 + 5x_2 = \phi(x_1) + \phi(x_2)$.

But $(x_1 \times x_2) = 5(x_1 \times x_2) = 5x_1 \times x_2 \neq \phi(x_1) \times \phi(x_2)$.

So ϕ is not a ring homomorphism.

(2) If $\phi(x) = x$

$\phi(x_1 + x_2) = \phi(x_1) + \phi(x_2)$

$\phi(x_1 \times x_2) = \phi(x_1) \times \phi(x_2)$

$\phi(1) = 1$

So $\phi(x) = x$ is a ring homomorphism.

And $\phi(x) = x$ is also a bijection, by **lemma 2.1** above since $\ker \phi = \{0\}$.

We call this function ϕ is a ring isomorphism.

- $\phi : Z \to R$

$\phi(x_Z) = x_R$ is also a ring homomorphism.

$\phi(x_Z) = x_R$ is injective., but not surjective, so this function ϕ is not a ring isomorphism.

- $\phi : C \to C$ the complex conjugation $\phi(z) = \overline{z}$ for $z \in C$ is also a ring isomorphism.
- $\phi : Z \to Z_m$ is a surjective homomorphism but not injective given by $\phi(a) = a(\text{mod } m)$, i.e., ϕ assigns to each integer the equivalence class of its remainder via division by *m*. I will later in use a particular section (Chapter 2.4) to talk about the modular equivalence relation.

From the above examples, we see Z or C is, of course, isomorphic to itself, while $Z \ncong Z_m$, $Z \ncong R$ (since there is no bijection existing between Z and Z_m (c.f., Chapter 2.4), Z and R (c.f., Chapter 5)). Moreover, all integers and all rational numbers are not "the same" structure of numbers either, though they both are "countable" (which I will talk about later in Chapter 5.3). Nevertheless, we all regard the complex number of the form (*x*, 0) and the real number *x* as being identical, because, in this case, we consider the real numbers as being embedded/included in the complex numbers by the embedding/inclusion map E: $x \to (x, 0)$ from R to C, and there between *x* (namely, the domain of mapping E) and (*x*, 0) (namely, the image of mapping E) is established an isomorphism. Later in Chapter 3 we will see another example of isomorphism between polynomials.

2.2 Logarithmic and exponential functions

Connections with concepts that students already know play an important role in engaging students in high-level thinking processes. Every high school student learns the exponential function in the form of $y = r^x$ with $r, x \in R$ and knows that it is the inverse

of a logarithmic function in the form of $y = \log_r(x)$ with $r \in R^{>0}$ and $x \in R$; and vice versa. In particularly, the exponential function with base e $y = e^x$, also written as $y = \exp(x)$, is the inverse of the natural logarithmic function $y = \ln(x)$; and vice versa. Students are generally told, at most, that, given domains and ranges, exponential functions or logarithmic functions are one-to-one and onto functions, or mappings, or correspondences. Students may also intuit this from the graphs of the functions, through which students can visualize algebraic concepts to some extent, and which I think is really important. Martin (2007) confirms that students "need opportunities to model concepts concretely and pictorially" (p. 35).

However, teachers have an opportunity to help students realize more about exponential functions and logarithmic functions than they are inverses of each other. Being inverses of each other means that each of them is a bijection, which is part of the conditions to be an isomorphism. Are exponential functions and logarithmic functions homomorphisms? An important feature of exponentials is that they reduce multiplication to addition, by the formula: $r^{x_1} \cdot r^{x_2} = r^{x_1 + x_2}$, which by taking logarithm with base r implies a corresponding important feature of logarithms that they also reduce multiplication to addition, by the formula: $\log_r{}^{(x_3 x_4)} = \log_r(x_3) + \log_r(x_4)$. Surprisingly, the two formulas just show they are group homomorphisms. Therefore, $r^{x_1 + x_2} = r^{x_1} \cdot r^{x_2}$ expresses a group isomorphism between the additive group of real numbers (denoted by $(R, +)$) and the multiplicative group of positive real numbers (denoted by $(R^{>0}, \cdot)$), while $\log_r{}^{(x_3 x_4)} = \log_r(x_3) + \log_r(x_4)$ expresses a group isomorphism between the multiplicative

group of positive real numbers (denoted by $\left(R^{>0},\cdot\right)$) and the additive group of real numbers (denoted by $\left(R,+\right)$). In other words, we say exponential functions are the continuous isomorphisms from the additive group of real numbers to the multiplicative group of positive real numbers; logarithmic functions are the continuous isomorphisms from the multiplicative group of positive real numbers to the additive group of real numbers. Symbolic notions for this would be that $\phi:(R,+)\to(R^{>0},\cdot)$ given by $\phi(x)=r^x$ is a group isomorphism; $\phi:(R^{>0},\cdot)\to(R,+)$ given by $\phi(x)=\log_r(x)$ is also a group isomorphism. In fact, in order to check they are isomorphisms, teachers may also guide students first to check they are group homomorphisms, and then use **Lemma 2.2** above to check they are injections and check they are surjections by definition.

In the above example, a group isomorphism connects two specific groups. To figure out what the isomorphism is and what the two groups are underscores the need for students to consider the operations, the domain and range of the function and to identify the meaning of the function's inverse for that specific isomorphism. Nevertheless, if students justify the meaningless nature of a function's inverse regarding the desired domain and range, they will realize that the function is not an isomorphism. Again, I suggest teachers pointing out these mathematical facts to students so that students will not be intimidated when they meet abstract algebra for the first time at colleges.

We all agree that meaningful contexts help students see the important features of a concept. Next I give two contextualized examples to show how useful exponential functions and logarithmic functions are as group homomorphisms. I will briefly mention

the application of them as bijections at the end of the second example because most students are acquainted with it.

If a certain principal (denoted by P) is invested at an annual rate (denoted by r in terms of decimal) compounded n times a year, then the amount (denoted by A) in the account at the end of t years is given by $A = P(1+\frac{r}{n})^{nt}$. Usually r and n are set by the bank, and the depositor can only make decision on t the number of years the principal is put in the bank. So if we let $M = P(1+\frac{r}{n})^{n}$ we have $A = M^{t}$, which can be considered as an exponential function with the base M, the independent variable $t \in N$, and the dependent variable A. Once we know the value of M we know the value of M^{t} for every $t \in N$. This specific functional relationship is the homomorphism property of exponentials because

$M^{t} = M^{1 \cdot t} = M^{\overbrace{1+1+\ldots+1}^{t \text{ 1's}}} = \overbrace{M^{1} \cdot M^{1} \cdot \ldots \cdot M^{1}}^{t \ M^{1}\text{'s}} = \overbrace{M \cdot M \cdot \ldots \cdot M}^{t \ M\text{'s}}$. If the annual rate is compounded continuously then $A = Pe^{rt}$. Due to the same homomorphism property, if we know e^{r} then we know $\left(e^{r}\right)^{t}$, and so we know $A = Pe^{rt} = P\left(e^{r}\right)^{t}$. Some student may understand the formula for compounding discretely but may not understand the origin of the formula for compounding continuously. Considering the formula of compounding discretely $A = P(1+\frac{r}{n})^{nt}$. Let $m = \frac{n}{r}$, then

$A = P(1+\frac{r}{n})^{nt} = P\left[(1+\frac{1}{m})^{m}\right]^{rt}$, where $(1+\frac{1}{m})^{m} \to e$ as $m \to \infty$ proved in Calculus. In

high school mathematics class teachers may guide students using spreadsheet to observe the values of $(1+\frac{1}{m})^m$ when m gets very large, or using some graphing software to draw the graph of the function $y=(1+\frac{1}{m})^m$ and then observe the trend of the graph when m gets very large, as shown in the following figure.

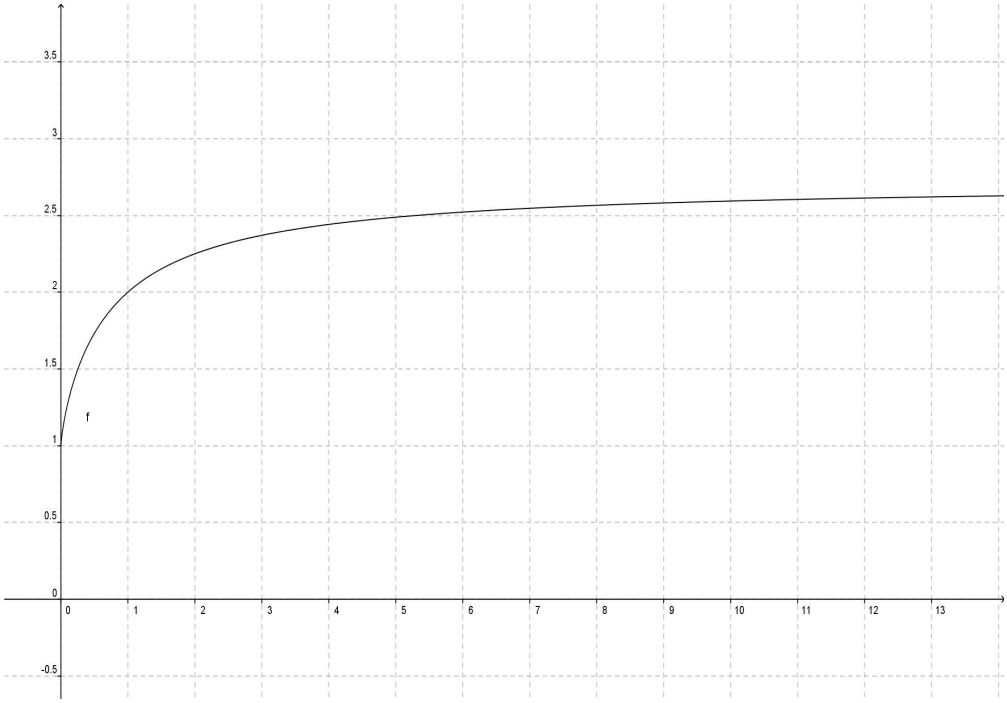


Figure 2.2.1 Partial view of the graph of the function $y=(1+\frac{1}{m})^m$

It is important for teachers to be cautious about using technology in the classroom by noting that all the values displayed by software are rounded to rational numbers, which never equal e because it is an irrational number. Nevertheless, the trend of the graph above shows that the greater m is, the greater y is (in Calculus we can prove $y=(1+\frac{1}{m})^m$ is an increasing function). So the greater n is, the greater m is and so A is. In other words, the greater the number of times the rate is compounded per year, the

greater the interest is if the principal, rate, and years are fixed. Thus the depositor would benefit more from the continuous compound interest, while some banks compound interest quarterly or monthly.

One advantage of logarithms is to make a large range of numbers manageable. The decibel is a logarithmic unit of measurement that expresses the magnitude of a physical quantity (such as intensity of sound) relative to a specified or implied reference level (usually in terms of ratio). When referring to measurements of amplitude, the decibel (dB) is defined by evaluating ten times the base-10 logarithm of the ratio of the squares of A_1 (measured amplitude) and A_0 (reference amplitude), i.e., $A_{\text{dB}} = 10\log(\frac{A_1^2}{A_0^2}) = 20\log(\frac{A_1}{A_0})$. When we want to calculate the overall decibel gain of the consecutive amplifiers (a multi-component system), we can simply compute the summation of the decibel gains of the individual amplifiers (components of the system), rather than multiplying amplification factors A_i/A_0. Essentially this is because of the homomorphism property of logarithms, i.e., $\log(\prod_i^n \frac{A_i}{A_0}) = \sum_i^n \log(\frac{A_i}{A_0})$. Hence, if we want the product of the large amplification factors, we can simply exponentiate the sum of the individual decibels that we have because exponential functions are inverses of logarithmic functions.

2.3 Conjugate zeros theorem

We say K is a **field extension** of F, if (K, F are fields) and (K contains F, i.e., $F \subset K$). We have already known that C and R are fields, and $R \subset C$, so C is a field extension of R. Similarly, C is a field extension of Q. And R is a field extension of Q.

In other words, we extend the field of rational numbers to the field of real numbers by adjoining irrational numbers, and extend the field of real numbers to the field of complex numbers by adjoining $i = \sqrt{-1}$. What if we do not adjoin all irrational numbers to rational numbers? In abstract algebra, for example, we can prove $Q[\sqrt{2}]$ is a field extension of Q by only adjoining $\sqrt{2}$ to Q (c.f., Shifrin, 1996, p. 54, Example 4), so it allows the smallest possible enlargement of Q (a number system) in which a given equation like $2x^2 - 1 = 0$ has roots. These are all numbers systems high school students are familiar with. Nonetheless, students just do not take them in this way, instead, they merely think Q, R, C are different set of numbers, and know one set contains the previous one.

Furthermore, in high school no student takes C as an R-vector space. Every high school student learns every element in C can be written uniquely as $a + bi$ with $a, b \in R$, but teachers hardly tell students $a + bi = a \cdot 1 + b \cdot i$, and so $\{1, i\}$ is a basis for C as an R–vector space, and then the degree of C over R, denoted by $[C : R]$, equals 2, which is the dimension of C as an R–vector space, i.e., the number of elements in the R-basis of C. This basic fact can be related to the following **proposition**:

Suppose K is a field extension of F and $\alpha \in K$ is the root of polynomial $f(x) \in F[x]$ which is irreducible in $F[x]$. Then $[F[\alpha] : F]$ = degree of $f(x)$.

Clarification: for instance, if we say a polynomial is in $R[x]$, we mean all the coefficients of the polynomial are all in R, i.e., the coefficients are all real numbers; for later use in Chapter 3, if we say a polynomial is not irreducible in $R[x]$, it means the polynomial cannot be factored into polynomials with coefficients all in R.

Let us still observe the example of $[C:R]$. Given $f(x)=x^2$-2x+3, the two roots of $f(x)$ are $1\pm\sqrt{2}i$, which are not in R, so $f(x)$ is irreducible in $R[x]$ (c.f., Chapter 5.3). C is a field extension of R, and $\alpha_1=1+\sqrt{2}i$ or $\alpha_2=1-\sqrt{2}i$ is contained in C. So $[R[\alpha_1]:R]=\deg(f(x))=2$ or $[R[\alpha_2]:R]=\deg(f(x))=2$. But $\sqrt{2}\in R$, so actually as mentioned above we only need to adjoin i to R to get C, i.e., $R[\alpha_1]=R[\alpha_2]=R[i]=C$, therefore, $[C:R]=\deg(f(x))=2$.

In high school, students are taught that for any polynomial in $R[x]$, non–real roots come in complex conjugate pairs (namely, the conjugate zeros theorem). What teachers usually do in class is telling students the fact and use specific examples to demonstrate it. I think this reality is decided by students' capability of understanding abstract mathematical concepts and by in-service teachers' understanding or mastering of how to roughly show, to the students especially those "enthusiasts", the idea of proving this mathematical fact using abstract algebra knowledge.

I will first try to state the related parts needed to prove the fact, and then relate it to the common proof given in high school or even in the first year of college.

Let K be a **field extension** of F with finite degree. Let $\phi: K\to K$ be a ring **isomorphism**. We say that the isomorphism ϕ is an F -**automorphism** of K if $\phi(a)=a$ for all $a\in F$, i.e., ϕ fixes the elements in F. Then the **Galois group** of K over F is defined as $G(K/F)$ = { F -automorphisms of K } with group operation (Note: here K/F doesn't mean a quotient ring, but just means K over F). Interestingly, we find the complex conjugation function is an R -automorphism of C. In other words, $G(C/R)$ is a Galois group, and complex conjugation is an element of the group, i.e., $\phi\in G(C/R)$.

Before we move on, let us look at a **lemma**:

K is a field extension of F. $\alpha \in K$ is a root of polynomial $f(x) \in F[X]$. For any $\phi \in G(K/F)$, $\phi(\alpha)$ is also a root of $f(x)$.

The proof of this lemma uses quite fundamental mathematical technique and the definition of automorphism we introduced above.

Proof.

Let $f(x) = C_0 + C_1 x + ... + C_n x^n$ for $C_0, C_1, ..., C_n \in F$.

$\alpha \in K$ is a root of $f(x) \Rightarrow C_0 + C_1\alpha + ... + C_n\alpha^n = 0$ (※).

Apply ϕ to (※). We get $\phi(C_0 + C_1\alpha + ... + C_n\alpha^n) = \phi(0)$.

Since ϕ is a ring homomorphism by definition, we have

$\phi(C_0) + \phi(C_1)\phi(\alpha) + \ldots + \phi(C_n)\phi(\alpha))^n = 0$.

And since ϕ is an F-automorphism of K, we get $C_0 + C_1\phi(\alpha) + \ldots C_n(\phi(\alpha))^n = 0$.

This expression just shows $\phi(\alpha)$ is a root of $f(x)$, as desired.

Now, by the lemma we just now proved, if $\alpha \in C$ is a root of a polynomial $f(x) \in R[x]$, then $\phi(\alpha)$, namely, the complex conjugation of α, is also a root of $f(x)$.

Therefore, we proved the mathematical fact "non-real roots of any polynomial in $R[x]$ come in pairs".

The usual proof that most high school students or college freshmen are given is, first, taking the conjugate of the whole equation (※), i.e., $\overline{C_0 + C_1\alpha + ... + C_n\alpha^n = 0}$, which is essentially the above step of applying ϕ to (※), and secondly, using the homomorphism property of ϕ that conjugates of sums and products are sums and products of conjugates

(essentially assuming $\phi \in G(K/F)$) to get $C_0 + C_1\overline{\alpha} + ... + C_n\overline{\alpha}^n = 0$ which means $\overline{\alpha}$ is a root of $f(x)$.

Dr. Roy Smith commented that the proof was too clever for him to think of as a freshman, and perhaps also for many other students, but maybe if they learn to think about symmetries they would appreciate the idea better; he suggested starting motivating students with the following fact: the sum of the vectors starting at the origin and ending at the vertices of a regular polygon of n sides with the symmetric center at the origin is zero (personal communication, February 9, 2010), because the compositions of the vectors on the horizontal axis and vertical axis are zero, and when we rotate the polygon through 360/n degrees or we flip it in the horizontal axis or the vertical axis, the vectors just exchange positions while the sum of them does not change.

Hence, next we use the idea of symmetry to rethink about the above proof of the conjugate zeros theorem. If there exists an imaginary number which we view as a vector starting at the origin and ending at the point which represents the number in the complex plane, then its square is another imaginary number corresponding to a vector, and so is its cube, and so forth till its highest power which equals the degree of the polynomial. And then we dilate these vectors by the corresponding coefficients of the polynomial and connect the terminal points of the new vectors, such that we have an (irregular) polygon, the sum of whose origin-starting vectors is zero. Thus this imaginary number is a root of the polynomial. Now concerning flipping the polygon in the real axis, those vectors turn into the vectors of the corresponding conjugates. The only thing it changes is the directions of these vectors, but the sum of the origin-starting vectors of the polygon is taken to its opposite value, “negative” 0, which means the sum 0 does not change. So the

complex conjugate is also a root of the polynomial. Again, all of these are about symmetry and isomorphism.

As an extension of this idea, every high school student should know that if a quadratic equation has real solutions, then the solutions come in pairs (including those with only one solution, which is in pair with itself). It is because the two solutions come symmetrically in the axis of symmetry of the graph of the quadratic function. So this is an analogue to the symmetry explanation of the conjugate zero theorem. Theoretically, the reflection in that axis of symmetry can be presumably considered as an automorphism as well.

Now let us go back to the example I mentioned at the beginning of the section that $Q[\sqrt{2}]$ is a field extension of Q. This is elegantly analogous to that C is a field extension of R by adjoining $i=\sqrt{-1}$, i.e., $C=R[\sqrt{-1}]$. If $\alpha\in Q[\sqrt{2}]$ with $a,b\in Q$ is a root of polynomial $f(x)\in Q[x]$, then $\phi(\alpha)\in Q[\sqrt{2}]$ should be another root of the polynomial, by the Q-automorphism of $Q[\sqrt{2}]$ defined as a $\phi(a+b\sqrt{2})=a+b\phi(\sqrt{2})$. For simplicity, we may also define ϕ as a "conjugation mapping" by $\phi(\sqrt{2})=-\sqrt{2}$, i.e., $\phi(a+b\sqrt{2})=a-b\sqrt{2}$. Let us check with, for example, the polynomial $x^2+2x-1\in Q[x]$. Let $x^2+2x-1=0$. We have $x=-1\pm\sqrt{2}\in Q[\sqrt{2}]$, as claimed the roots coming in conjugate pairs. What about the polynomial $x^2-3\sqrt{2}x-(1+3\sqrt{2})$? Obviously it is not in $Q[x]$, so the roots coming in conjugate pairs is not guaranteed. To be sure we check the roots of $x^2-3\sqrt{2}x-(1+3\sqrt{2})$. They are $1+3\sqrt{2}$ and -1, no conjugate pairs. Moreover, if we are comfortable with notations we can let $\sqrt{2}$ be any symbol we like, say $\sqrt{2}=\tau$.

Thus the field extension $Q[\tau]$ of Q would look more like $C = R[i]$ as a field extension of R. And as having the complex plane, we may have our own "$Q-\tau$ plane" as we wish. If we adjoin all square roots of rational numbers to Q, we will have the Euclidean plane (more about which we will see later in Chapter 5.4). Please note that I just presented an analogy between $Q[\sqrt{2}]$ and $R[\sqrt{-1}]$, but I did not say there is an isomorphism between them. Actually it is impossible to establish one here, because Q is countable while R is not (we will talk about this later in Chapter 5.3). What about $Q[\sqrt{2}]$ and $Q[\sqrt{-1}]$? They both are countable but still not isomorphic to each other because $Q[\sqrt{-1}]$ contains $\sqrt{-1}$ and so contains square root of negative rational numbers whereas $Q[\sqrt{2}]$ does not. Hence, this should be the more germane reason for why $R[\sqrt{-1}]$ is not isomorphic to $Q[\sqrt{2}]$.

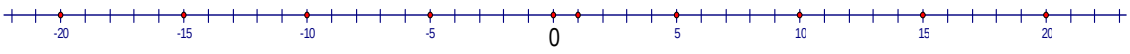
It is not easy for students to reach the kind of understanding that we have talked about in this section unless they have a really good understanding of the mathematical definitions. In addition, after the students are exposed to examples of one type the teacher, who has much more experience and a better understanding of mathematics, needs to help students see the connections I have described. Hence, this is why I concentrate on the connections themselves and always refer to the related basic definitions.

2.4 Modular equivalence relation

When we do long division, we always divide an integer by a smaller one, and the quotient number is always as big as possible until the non-negative remainder is smaller than the dividing number. Mathematicians summarized this rule into the **Division Algorithm for natural numbers**:

Given a, b$\in N$, there are integers q (for "quotient") and r (for "remainder") so that a=q×b + r, with 0≤r≤b.

In our experience, when we divide different a's by the same b, consequently q's are different, but r could be the same. This is an interesting and useful phenomenon. According to Cuoco (1990, p. 264), the word "modulo" used by Gauss is from the Latin verb that means "to measure". Roughly speaking, if we use 5 as a measure on the number line



then 17 is 2 more than a marked number 10. So Gauss would say that "17 is 2 modulo 5", or "17 is 2 more than a multiple of 5", or "17 is 2, except for a multiple of 5". All these can be expressed symbolically as $17 \equiv 2 \pmod{5}$. Cuoco (1990, p. 264) commented that the custom of thinking of "modulo" as "except for" has added a rich flexibility to the word, and then he illustrated the beauty of "modulo" by telling us that mathematicians often say things like "I can prove this theorem modulo one conjecture" or even "Modulo a rainstorm, we will have a picnic this afternoon".

People notice this kind of phenomenon appears in our everyday life and is involved in some basic calculation, and call it "modular arithmetic". I include several examples below.

(1) For example, 3 o'clock in the morning is 3 past the position 12:00; 15 o'clock in the afternoon is also 3 past the position 12:00 of the clock. We symbolize this as $3 \equiv 15 \pmod{12}$. This illustrates why the modular arithmetic sometimes is also called clock arithmetic.

(2) Interestingly, when people call different days "Monday" they actually use the idea of "mod 7", so we may number the seven days in a week as {Mon.=1, Tue.=2, Wed.=3, Thu.=4, Fri.=5, Sat.=6, Sun.=0}, and when it moves from one day to next day, we algebraically add 1 to any number and it will become the next number, if we define 6+1=7=0 in this context.

(3) Suppose Adam has 93 cents and Bob has 38 cents. They want to convert the cents they have into as many nickels as possible and see who has more cents left. Because 93=18×5＋3. Adam gets 18 nickels and 3 cents. Similarly, Bob gets 7 nickels and 3 cents. So they have the same amount of cents left. Symbolizing this is $93 \equiv 38 \pmod{5}$.

(4) In trigonometry $\sin 30^\circ = \sin 390^\circ$ because $30^\circ \equiv 390^\circ \pmod{360^\circ}$.

(5) We know $i^2 = -1$ so $i^4 = 1$, then we can compute, for examples, $i^7 = i^3 = i^{2+1} = -i$ and $i^{7348} = i^0 = 1$.

(6) Actually, our decimal division uses 10 as the dividing number. If integers a_1 and a_2 divided by 10 have the same remainders, then we symbolize that $a_1 \equiv a_2 \pmod{10}$.

(7) I remember a brilliant explanation about $5 \equiv 0 \pmod{5}$ generated by a homeschooled Algebra I boy (L. Holladay, personal communication, January 18, 2010) as follows:

In a chemical equation, the mass of all the reactant substances must be equal to the mass of all the product substances. If the combined mass of all the reactant substances in a reaction is five grams, the product mass will also be five grams. If you are looking for the mass of the product which you can contain, the masses may be different. If the reaction caused substances with a combined mass of five grams to turn into gases which you are unable to contain, the mass of product would be zero.

In teaching students it is good to start with a context and extract the mathematics and then later on apply it to new situations. Thus, let us formally define "integer arithmetic modulo *m*" as follows: If $a, b \in Z$ are "equal after casting out *m*'s", i.e., $a \equiv r \pmod{m}$ $\Leftrightarrow a = q_1 \cdot m + r$, $b \equiv r \pmod{m} \Leftrightarrow b = q_2 \cdot m + r$, where *r* is the remainder, q_1, q_2 are the quotients by division algorithm, then we write $a \equiv b \pmod{m}$ (read "a is congruent/equivalent to b mod *m*"). It is easy to check this is an equivalence relation on Z, and equivalence modulo *m* respects the algebraic operation on Z, i.e., **Proposition 3.1** (Shifrin, 1996, p. 21):

If $a_1 \equiv b_1 \pmod{m}$ and $a_2 \equiv b_2 \pmod{m}$, then (1) $a_1 + a_2 \equiv b_1 + b_2 \pmod{m}$; (2) $c \cdot a_1 \equiv c \cdot b_1 \pmod{m}$; and, more generally (3) $a_1 \cdot a_2 \equiv b_1 \cdot b_2 \pmod{m}$.

A special case is $m \mid a \Leftrightarrow a \equiv 0 \pmod{m}$. So we rewrite $a \equiv b \pmod{m}$ into $a - b \equiv 0 \pmod{m}$, therefore, $a \equiv b \pmod{m} \Leftrightarrow m \mid (a - b)$.

If an integer is divided by *m* then all the possible remainders are {0, 1, 2, …, *m*-1}. For each remainder *r*, there is a set of integers corresponding to it, such that every element *n* in the set divided by *m* has remainder *r*, i.e., $n \equiv r \pmod{m}$ and all such *n*'s

form the equivalence class of corresponding *r*. In other words, the integers Z is divided into *m* equivalence classes of each possible remainder upon the division by *m*. Symbolically,

$\{ n \in Z : \ n \equiv r \pmod{m}, \ r = 0, 1, 2, ..., m\text{-}1 \}$

$= \{ \{ ..., -3m, -2m, -m, 0, m, 2m, 3m, ... \},$

$\{ \ldots, -3m+1, -2m+1, -m+1, 1, m+1, 2m+1, 3m+1 ... \},$

$\{ ..., -3m+2, -2m+2, -m+2, 2, m+2, 2m+2, 3m+2, ... \},$

$\ldots,$

$\{ \ldots, -3m+(m-1), -2m+(m-1), -m+(m-1), m-1, m+(m-1), 2m+(m-1), 3m$

$+(m-1), ... \} \}$

$= mZ + r$.

And we denote the finite modular ring by $Z/mZ = Z/\langle m \rangle = Z_m = \{\bar{0}, \bar{1}, \ ... \ , \overline{m\text{-}1}\}$. If there is no confusion, sometimes we omit the bar " $\bar{}$ " on the number modulo *m*. Hence,

$Z_m = \{0, 1, 2, ..., m\text{-}1\}$.

It is obvious that there is a 1-1 correspondence between each equivalence class of the remainder modulo *m* and the remainder itself. However, as we mentioned before in Chapter 2.1, $\phi : Z \to Z_m$ defined by $\phi(a) = a \pmod{m}$ is not a 1-1 correspondence, since apparently Z is infinite and Z_m or $Z/\langle m \rangle$ is finite and it is impossible to have a bijection between an infinite set and a finite one.

Nevertheless, luckily, $\phi : Z \to Z_m$ is a homomorphism, as a consequence of the above **Proposition 3.1** (good exercise for students). Generally, a homomorphism $\phi : A \to B$

carries over structure partially or fully from one setting A to another B. In other words, the operation in A is preserved in B by the homomorphism ϕ, and if ϕ is a bijection then any structure in A is completely preserved in B, and vice versa. Losing some of the structure or simplifying structure while still preserving the operation(s), as a homomorphism does, can make problem solving sometimes easier whereas an isomorphism does not. So it is key to being able to deal with the algebraic structure preserving feature of a function, which sometimes is even more basic than the bijection feature.

Now let us concentrate on the homomorphism $\phi: Z \to Z_m$ defined by $\phi(a) = a \pmod{m}$. It is obvious that some structure of Z is preserved in Z_m while some other is lost. As noted above the actual number of elements is no more infinite for Z_m than it is for Z, but an integer equation in Z_m still has integer solution(s) if it is solvable in Z. Because applying modulo m is a homomorphism, we conclude that if a statement is true, then it is also true for every (integer) modulus m. Its equivalent contrapositive says that if a statement does not hold for every (integer) modulus m, then it is also not true before we apply modulo m to it. To understand this property better and show what modular arithmetic is good for, we take a look at the following little nice techniques.

(1) We can check $378+295 \neq 674$ by only checking the "ones digits" of those three items without adding them together. The contradiction is that $378 \equiv 8 \pmod{10}$, $295 \equiv 5 \pmod{10}$, $674 \equiv 4 \pmod{10}$. But 8+5=13, $13 \equiv 3 \pmod{10}$, $3 \not\equiv 4 \pmod{10}$.

(2) We can generalize the Divisibility Criteria (c.f., Proposition 3.2 (Shifrin, 1996, p. 21)), which most of the high school students are familiar with.

(3) We can test for divisibility by 11. $10 \equiv -1 (\text{mod } 11)$ and $100 \equiv 1 (\text{mod} 11)$ can be written in forms of $\overline{10} = \overline{-1}$ in Z_{11} and $\overline{100} = \bar{1}$ in Z_{11}. Thus, $\overline{1000} = \overline{100} \cdot \overline{10} = \overline{-1}$ in Z_{11}, $\overline{10000} = \overline{1000} \cdot \overline{10} = \overline{-1} \cdot \overline{-1} = \bar{1}$ in Z_{11}, and so forth. For example, $\overline{6836542}$ = $\bar{2} - \bar{4} + \bar{5} - \bar{6} + \bar{3} - \bar{8} + \bar{6}$ = $\overline{2-4+5-6+3-8+6}$ = $\overline{-2} \neq \bar{0}$ in Z_{11}. Hence, 6836542 is not divisible by 11.

(4) The solutions to some classic problems like "telling the day of the week for any date" and "casting out nines" are also based on modular arithmetic.

(5) We can use, for example, arithmetic mod 2 to test for the solvability of some integer equations, such as $34x^4 + 16x^3 - 8x - 103 = 0$. If we subtract the constant term -103 from the equation, we will get the transformation $34x^4 + 16x^3 - 8x = 103$. We notice that every term on the left-hand-side of the equation is even equivalent to 0 (mod 2), but the right-hand-side is odd equivalent to 1 (mod 2). Thus to solve this equation for integers is impossible.

(6) Arithmetic mod 4 can be used to detect perfect squares. Any perfect square must be equivalent to either 0 or 1 (mod 4), because $0^2 \equiv 0 \text{ (mod 4)}$, $1^2 \equiv 1 \text{ (mod 4)}$, $2^2 \equiv 0 \text{ (mod 4)}$, $3^2 \equiv (-1)^2 \equiv 1 \text{ (mod 4)}$. So no matter how big the integer 983747823194209432612542015 is, it is equivalent to 3 (mod 4) (noticing the last two digits of the number). So it is safe to conclude that the given integer is not a perfect square.

(7) Then, let us try to solve equations like $x^2 + y^2 = n$ where $x, y \in Z$ and $n \in N$. Reducing the equation by mod 4, according to the possible values (mod 4) of a

perfect square, x^2+y^2 can only be 0, 1, or 2. So if $n \equiv 3 \pmod 4$ then $x^2+y^2=n$ cannot be solved for integers x and y. Moreover, suppose n is prime. Hence, if $n \equiv 1 \pmod 4$ then the equation is solvable for integers. This is consistent with Fermat's Last Theorem (proved using very deep methods by Andrew Wiles in 1995).

(8) We can solve the equation $x^3=1$ for $x \in C$.

$x=1^{\frac{1}{3}}=\left(\cos(2k\pi)+i\sin(2k\pi)\right)^{\frac{1}{3}}=\cos\left(\frac{2k\pi}{3}\right)+i\cdot\sin\left(\frac{2k\pi}{3}\right)$, or $x=1^{\frac{1}{3}}=\left(e^{i2k\pi}\right)^{\frac{1}{3}}=e^{i\left(\frac{2k\pi}{3}\right)}$, where $k \in Z$.

$k \equiv 0 \pmod 3 \Rightarrow x=1;$

$k \equiv 1 \pmod 3 \Rightarrow x=\cos\frac{2\pi}{3}+i\cdot\sin\frac{2\pi}{3}=-\frac{1}{2}+i\cdot\frac{\sqrt{3}}{2};$

$k \equiv 2 \pmod 3 \Rightarrow x=\cos\frac{4\pi}{3}+i\cdot\sin\frac{4\pi}{3}=-\frac{1}{2}-i\cdot\frac{\sqrt{3}}{2}.$

In other words, $k \pmod 3$ corresponds to one of the three complex cube roots of 1. The following is the picture illustration of this example.

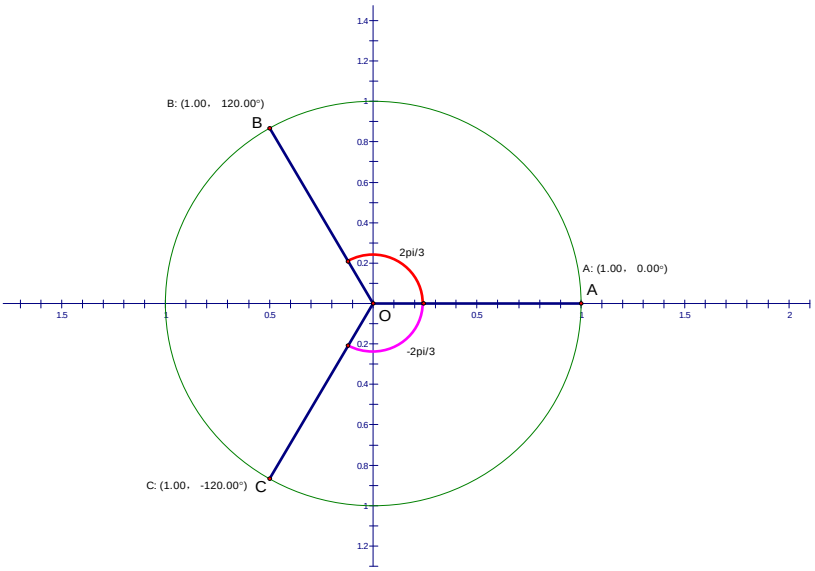


Figure 2.4.1 Illustration of the three complex cube roots of 1

(9) We will see at the end of the next chapter (Chapter 3) an interesting and more

realistic application of modular arithmetic based on Fermat's Little Theorem.

(10) Note that if an integer equation has no solution in Z it does not mean it has no solution in Z_m. For example, $x^2 - 2 = 0$ has no integer solution in Z, whereas $x^2 - 2 = 0$ may have solutions in Z_m. Suppose $m = 7$, then $x = 3$ satisfies the equation in Z_7. This is because that a homomorphism partially preserves the structure such as solvability, but it does not take care of the non-existent structure. In other words, a homomorphism takes solutions to solutions but does not take non-solutions to non-solutions. Dr. Smith supplemented that since some things are taken to zero, the error that measures how far an integer is from being a solution may be taken to zero and then the integer becomes a solution (personal communication, April 13, 2010).

Note that the algebraic expressions are powerful to represent mathematical relationships in the given contexts, embody people's insights into these relationships, and help people record ideas and organize thoughts mathematically.

One way to connect the idea of modular arithmetic to the experiences of younger students is through arithmetic computation "tricks" that students often learn. For example, some students were taught in elementary school an interesting method of multiplying two numbers between 5 and 10 with the aid of their fingers. To calculate 6×7, one needs to raise (6-5=1) finger on one hand and (7-5=2) fingers on the other, and then add the numbers denoted by the raised fingers, i.e., 1+2=3, and multiply those denoted by the bent fingers, i.e., (5-1)×(5-2)=4×3=12. Thus, the product is (3×10+12=42). Some young students might be happy to learn this trick, especially when they learn arithmetic. However, they seldom ask why this trick works, even after they become high school

students knowing more mathematics, though they have the knowledge and ability to figure out why it works.

Verifying this trick is within the mathematical expertise of high school students and makes use of ideas from modulo equivalence classes. In order to create a justification, we need to model the situation; usually the first step of modeling is to define the variables. In this instance, let $a,b \in [5.10]$ be the two integers which we want to multiply, and then a-5, b-5 , respectively, will denote the number of raised finger(s) on each hand. You may make a different decision to define the variables (for example, let $a,\ b$ be the number of raised finger(s) on each hand); it is up to you as long as your model works for the rule(s).

Thus, we want to check whether $a \cdot b$ equals $[(a-5)+(b-5)] \cdot 10+[5-(a-5)] \cdot [5-(b-5)]$. Simplifying the latter does yield the former $a \cdot b$, so they are equal; that is why the trick works.

This trick actually entails some abstract algebra knowledge. In detail, when we calculate the number of raised bent fingers, a-5 in Z (not a field) is $\overline{a}$ in field Z_5, and then $5-(a-5)$ in Z is $(\overline{a}^{-1}) = (5-\overline{a}) = (0-\overline{a}) = (-\overline{a})$ in Z_5; similar for b. But $(\overline{a}^{-1})(\overline{b}^{-1}) = [5-(a-5)] \cdot [5-(b-5)]$ in Z_5 may not be equal to the value itself any more in Z, since Z is not a field ; the calculation rules in Z do not work in the world consist of some elements of Z, for example $Z_5 = \{0,1,\ldots 4\}$. That is why, in the case 6×7, the additional inverse $\overline{x}^{-1}$ of $\overline{x} \in Z_5$ is $(0-\overline{x} = -\overline{x})$ or $(\overline{5}-\overline{x})$, so $(\overline{6}^{-1})(\overline{7}^{-1})$ $= (\overline{1}^{-1})(\overline{2}^{-1}) = (-\overline{1})(-\overline{2}) = \overline{2}$ in Z_5, but also $(\overline{6}^{-1})(\overline{7}^{-1}) = (\overline{1}^{-1})(\overline{2}^{-1}) = \overline{4} \cdot \overline{3}$ in Z_5, while then $4 \cdot 3 = 12$ in Z. $2 \neq 12$ in Z. Similarly, $\overline{a}+\overline{b}$ in Z_5 may not be equal to $\overline{a}+\overline{b}$ in Z.

However, I think this kind of rule may confuse students. Some students may feel it makes the computation more complicated, and it does not seem to arise intuitively from the usual ways to compute, neither 6×7=(5+1)(5+2) $= 5^2 + 5(1+2) + 1 \cdot 2$ nor 6×7= (10-4)(10-3)= 10^2-(4+3)×10+4×3. Furthermore, it will be more confusing if the teacher shows students the case for example, 7×8, and tells them that 7-5=2, 8-5=3, and then 2+3=5 "is the tens, 50", and 3×2=6 "is the units", "the product being 56" (Smith, 1958, p. 201, as cited in Arcavi, 2008, p. 43). In this example, the additional inverse of $\bar{a}$ in Z_5 happens to be $\bar{b}$ which is different from $\bar{a}$, and the product of the inverses of $\bar{a}$ and $\bar{b}$ happens to be a one-digit positive integer, which can be called "the units" but what if it is two-digit positive integer, like in the example 6×7? Students may wonder whether they need to "take off" the "1" in "12" to only add the units "2" to the "tens" "30". Hence, teachers need to be careful about the choice of numbers and the wording of the "trick."

To end this chapter I will only introduce the following theorem and examples to point to a little bit higher level which student will eventually be able to understand and deal with, and hope to fire some students' interest in further pursuit of mathematics.

Referring to the **Fundamental Group Homomorphism Theorem** (Shifrin, 1996, p. 194), we can establish $\phi: Z_{12} \to Z_3$ (a map from an additive abelian group to another) $\phi([a]_{12}) = [a]_3$ as a subjective homomorphism, and the kernel of ϕ is $\langle [3]_{12} \rangle$, i.e., $\ker\phi$ is generated by 3(mod 12), denoted by $\ker\phi = \langle [3]_{12} \rangle$ = {{0, 3, 6, 9} in Z_{12} }= $3Z_{12}$, so $Z_{12} / \ker\phi \cong Z_3$.

Here are two classic examples.

(1) Groups $G=(R,+)$, $H=(Z,+)$, $G'=(\{z\in C:|z|=1,\cdot)$. $\phi: G\to G'$ given by $\phi(x)=e^{i2\pi x}=\cos(2\pi x)+i\sin(2\pi x)$ is a surjective homomorphism and $\ker\phi=\mathrm{Z}$ (the groups G/H and G' both contain the multiplicative identity 1), so $G/H\cong G'$ or $R/Z\cong\{z\in C:|z|=1\}$. Dr. Roy Smith offered the following metaphor (personal communication, December 27, 2009). When we play with a yo-yo, we wrap the string around the round yo-yo. If we consider the string with the directions of wrapping to be real numbers, consider the round yo-yo to be the circle of complex numbers of length one, and consider the number of times the string goes around with the directions of wrapping to be an integer. It will help us to understand the one-to-one correspondence between the elements of R/Z and those of $\{z\in C:|z|=1\}$. In other words, real numbers on the real line turn out to be real values of angles with center at the origin (0, 0), and each angle $\theta\in[0,2\pi)$ where $\theta=2\pi x$ while $x\in[0,1)$ is mapped one-to-one and onto each complex value of the circle of length one $e^{i\theta}$. Interestingly, the quotient of two additive groups turns into a multiplicative group, i.e., addition of angles in R changes into the definition of multiplication for complex numbers of length one $\{z\in C:|z|=1\}$. Algebraically it is just exponentiation, $e^{i2\pi s+i2\pi t}=e^{i2\pi s}\cdot e^{i2\pi t}$ with $s,t\in R$.

(2) Groups $C^{\times}=(C-\{0\},\times)$, $S=(R^{>0},+)$, $H=(\{z\in C:|z|=1,\times)=$ $(\{z\in C: z=e^{2\pi it},\times)$.

$\phi: C^{\times}\to S$ by $\phi(z)=|z|$ is a surjective homomorphism. $\ker\phi=H$. So $C^{\times}/H\cong S$. So to speak, the unit complex circle corresponds to 1, and other equivalent circles correspond to the positive non-zero real numbers, illustrated in the picture below. It might be helpful

if you connect the mapping modulo $H = (\{z \in C :| z |= 1, \times)$ to the closing of the open train of a peacock. Conceive of the closed train as the positive *x*-axis except for 0, more precisely $S = (R^{>0}, +)$ and the fully opened train as the complex plane except for 0, more precisely $C^{\times} = (C - \{0\}, \times)$. Also interestingly, the quotient group of two multiplicative groups turns into an additive group, i.e., the multiplication in $C - \{0\}$ changes into the addition of angles in $R^{>0}$. Algebraically it is exponentiation, $r_1 e^{i2\pi s} \cdot r_2 e^{i2\pi t} = r_1 r_2 e^{i2\pi s + i2\pi t}$ with $r_1, r_2, s, t \in R^{>0}$.

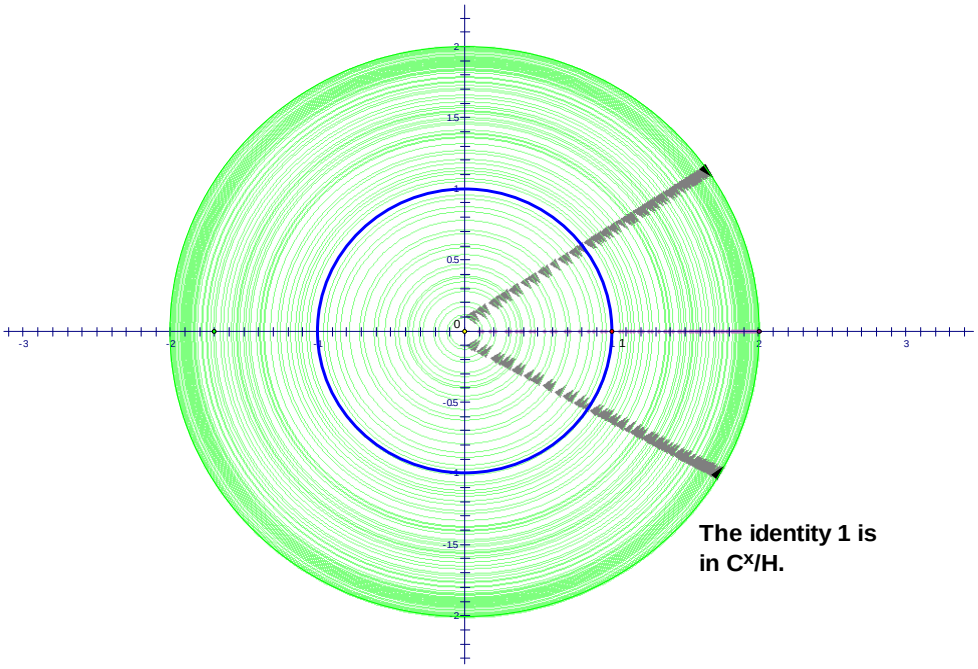


Figure 2.4.2 Illustration of the isomorphism $C^{\times} / H \cong S$

3. Binomial Coefficient and Irreducibility of Polynomials

In this chapter I first summarize several ways to develop the coefficient formula $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ of Binomial Theorem, and several places to use the formula to develop some other conclusions. Then I focus on the proof that the polynomial $x^{p-1} + x^{p-2} + ... + x + 1$ (p prime) is irreducible in $Q[x]$, which is related to Eisenstein's Criterion (our first irreducibility criterion; the other two in Chapter 5.3) and again the notion of isomorphism. Then I use one of the introduced lemmas and mathematical induction to give a brief proof and a realistic application of Fermat's Little Theorem, which involves no more mathematical knowledge than high school algebra and modular arithmetic.

Now let us start listing some connections that can be explored related to the binomial coefficient.

Briefly, the Binomial Theorem, i.e., $(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^{n-k} b^k$ for any $n \in N$, can be developed by exploding an n-dimensional solid with side of length $a+b$ such that students do not have to memorize it without understanding. Particularly, in spite of verifying the coefficient formula, namely $\binom{n}{k} = \frac{n!}{k!(n-k)!}$, by mathematical induction (c.f., Shifrin, 1996, p. 7), it can be generated by choosing a k-element subset of an n-

element set (c.f., Shifrin, 1996, p. 6) or by employing **group action** (c.f., Shifrin, 1996, p. 218) – I will introduce the concept of group action later in Chapter 4.3.

Conversely, the Binomial Theorem and its coefficient formula can be used to

(1) count the number of elements of the symmetric group S_5 which are conjugate (defined in Shifrin, 1996, p. 192) to the element (123) (Hint: $p_5^3 = \frac{5!}{3!} = 20$, or $C_5^3 \cdot 2 = \binom{5}{3} \cdot 2 = 10 \cdot 2 = 20$ since we first do not consider the order of the three entries of a 3-cycle, and then notice that a cycle and its inverse are different.);

(2) count the possibilities of some configuration when using Burnside's Theorem. The use of Burnside's Theorem is quite related to our life and other subjects. For example, we can find the number of different types of circular necklaces that can be made from six white and four blue beads. We can answer the question "how many different chemical compounds can be made by attaching H, CH_3, C_2H_5, or Cl radicals to the four bounds of a carbon atom?" We can also figure out how many ways one can paint the outer faces of a cube with several different colors (assuming we paint one whole face with only one color);

(3) prove that the polynomial $x^{p-1} + x^{p-2} + ... + x + 1$ (p prime) is irreducible in $Q[x]$ (this fact is an ingredient of the proof that the Galois group of a polynomial $f(x)=x^p - 1$ (p prime) is cyclic (Shifrin, 1996, p. 281, Exercise 23)).

Here, I want to spread out the proof that the polynomial $x^{p-1} + x^{p-2} + ... + x + 1$ (p prime) is irreducible in $Q[x]$ (*). First, the test for irreducibility of polynomials with integer coefficients lays the foundation for the deeper study of polynomials. Second, this proof

itself involves a basic irreducibility criterion of polynomials, and can be completed using students' understanding of school algebra. Consistent with the NCTM (2000) Standards' direction that instruction should enable all students to recognize reasoning and proof as fundamental aspects of mathematics, I approach this proof in a manner accessible to high school students.

Every big problem in the world grows from one or more small problems. If we are able to tear a big problem apart into sub-problems we may see how to solve the big one. This "subgoal" strategy is suggested in a lot of literature and research on cognition. Below I show how to break this larger proof into smaller pieces and identify some "nuggets" for high school students following each of the three lemmas below.

Before proving (*) we introduce **Eisenstein's Criterion** (c.f., Shifrin, 1996, p. 109, Theorem 3.5) as our **first irreducibility criterion** (we will see the other two in Chapter 5.3):

Given polynomial $f(x)=a_0+a_1x+a_2x^2+...+a_nx^n$, and $f(x)\in Z[x]$, which means the coefficients of $f(x)$ are all in Z. $f(x)$ is irreducible in $Q[x]$ if $\exists p\in Z(p \text{ is prime})$ such that $p|a_0$, $p|a_1$, … , $p|a_{n-1}$, but $p\not|a_n$, $p^2\not| a_0$.

For instance, $15x^5-10x^3+8x+14$ is irreducible in $Q[x]$ by Eisenstein's Criterion because we may test it with $p=2$; $8x^3-27x^2-6$ is irreducible in $Q[x]$ because we may apply Eisenstein's Criterion to it with $p=3$. The polynomial $f(x)=x^2+x+1$ seems to be impervious to Eisenstein's Criterion, but if we consider $f(x+1)=(x+1)^2+(x+1)+1$ is irreducible in $Q[x]$ due to Eisenstein's Criterion, then we can conclude that $f(x)=x^2+x+1$ is also irreducible in $Q[x]$. The upcoming proof explains the reason.

Note that I once made a low-class mistake of using “if and only if” in this statement instead of the correct “if (but not only if)”. It is apparent that there are polynomials $f(x) \in Z[x]$ irreducible in $Q[x]$ that do not satisfy Eisenstein’s Criterion, such as x^2+2 and x^3-32 (because of Root-Factor Theorem, which will be talked in Chapter 5.3). Thus this may be used as a good chance for the teacher to explain to the students the difference between a necessary condition and a sufficient one.

Then we need to prove three lemmas as follows.

Lemma 1. Any polynomial g(x+1) is irreducible in a field F[x] $\Rightarrow$ polynomial g(x) is irreducible in F[x].

Proof. We want to show g(x) is not irreducible $\Rightarrow$ g(x+1) is not irreducible, i.e., g(x) is reducible $\Rightarrow$ g(x+1) is reducible in F[x]. g(x) is reducible means g(x)=h(x)k(x) where h(x), k(x)$\in$ F[x]. We substitute x+1 for x in polynomial g(x) to get g(x+1)=h(x+1)k(x+1), which means g(x+1) is reducible in F[x], as desired.

Nuggets: definition of reducibility, proving contrapositive, replacing variables.

Lemma 2. $(x+1)^{p-1}+(x+1)^{p-2}+\ldots+(x+1)+1 = \dfrac{(x+1)^p-1}{x}$.

Proof. $x^p-1=(x-1)(1+x+x^2+\ldots+x^{p-1}) \Rightarrow 1+x+x^2+\ldots+x^{p-1}=\dfrac{x^p-1}{x-1}$ (※)

(Depending on the context we are using, we may or may not need the "as long as x does not equal 1" part. – I will offer a reason for this later in Chapter 5.3.).

Substituting x+1 for x in $1+x+x^2+\ldots+x^{p-1}=\dfrac{x^p-1}{x-1}$, we get

$(x+1)^{p-1}+(x+1)^{p-2}+\ldots+(x+1)+1=\dfrac{(x+1)^p-1}{x}$, as desired.

Nuggets: factoring and dividing a polynomial, replacing variables.

We now use (※) to show 0.999…=1, about which I will talk more later in Chapter 5.2.

Proof.

We use the idea of geometric series to turn a repeating decimal into the summation of fractions.

$0.999\ldots = \frac{9}{10}+\frac{9}{100}+\frac{9}{1000}+\ldots = \frac{9}{10}+\frac{9}{10^2}+\frac{9}{10^3}+\ldots = \frac{9}{10}(1+\frac{1}{10}+\frac{1}{10^2}+\frac{1}{10^3}+\ldots) =$

$\frac{9}{10}\left(\sum_{i=1}\left(\frac{1}{10}\right)^i\right)$, by (※), which equals $\frac{9}{10}\cdot\frac{\left(\frac{1}{10}\right)^{n+1}-1}{\frac{1}{10}-1}$ as n approaches infinity. But

$\left(\frac{1}{10}\right)^{n+1} \to 0$ when $n \to \infty$, so $\frac{9}{10}\cdot\frac{\left(\frac{1}{10}\right)^{n+1}-1}{\frac{1}{10}-1} = \frac{9}{10}\cdot\frac{0-1}{-\frac{9}{10}} = 1$ when $n \to \infty$. Therefore, 0.999…=1, as desired.

Lemma 3. $p\left|\binom{p}{k}\right.$ (p prime) for k=1, 2, …, p-1.

Proof. $\binom{p}{k} = \frac{p!}{k!(p-k)!} = p\cdot\frac{(p-1)!}{k!(p-k)!} \Rightarrow p\cdot(p-1)! = \binom{p}{k}\cdot(k!(p-k)!)$. $p\left|(p\cdot(p-1)!)\right.$ but $p \not| (k!(p-k)!)$ since p is prime. So $p\left|\binom{p}{k}\right.$.

Nuggets: transformation of combination formula (namely, the binomial coefficient), properties of prime numbers, the equivalence relation indicated by the equal sign "=" (which students usually do not understand well).

Now we prove (*) as follows.

Proof. The coefficient of x^k, namely a_k, in the expansion of $(x+1)^p$ is $\binom{p}{k}$, with k = 0, 1, . . . , p. So the coefficient of x^k in the expansion of $\frac{(x+1)^p-1}{x}$ is $\binom{p}{k+1}$, with k = 0, . . . , p-1. So by **lemma 2** the coefficient of x^k in the expansion of $(x+1)^{p-1}+(x+1)^{p-2}+...+(x+1)+1$ is $\binom{p}{k+1}$, with k = 0, . . . , p-1. Then by **lemma 3** $p\mid\binom{p}{k+1}$ with k = 0, 1, . . . , p-2. In the expansion of $(x+1)^{p-1}+(x+1)^{p-2}+...+(x+1)+1$, when k=p-1, $a_k=\binom{p-1}{p-1}=1$, so $p\nmid 1 \Rightarrow p\nmid a_{p-1}$; when k=0, $a_k=\binom{p}{0+1}=p$, so $p^2\nmid p \Rightarrow p^2\nmid a_0$. By **Eisenstein's Criterion** $(x+1)^{p-1}+(x+1)^{p-2}+...+(x+1)+1$ is irreducible in $Q[x]$. Thus, by **Lemma 1** $x^{p-1}+x^{p-2}+...+x+1$ is irreducible in $Q[x]$. Done. Thus, now we can safely say that $x^4+x^3+x^2+x+1$ is also irreducible in $Q[x]$.

Writing down the steps and writing about the relevant ideas helps students clarify their thinking and develop understanding.

Here is a question that I think needs our attention and consideration: How do we realize that in order to prove $f(x)=x^{p-1}+x^{p-2}+...+x+1$ (p prime) is irreducible in $Q[x]$ we should prove $f(x+1)=(x+1)^{p-1}+(x+1)^{p-2}+...+(x+1)+1$ (p prime) is irreducible in $Q[x]$?

I do not know the history of this "trick" and can only guess how it was discovered. If $f(x)$ can be factored, then $f(x+n)$ can be factored. If $f(x+n)$ can be factored then $f(x)$ can be factored, because $f(x) = f(x+n-n)$. Thus $f(x)$ is irreducible if and only if $f(x+n)$ is irreducible. So we look for a convenient value of the integer n. To apply Eisenstein's

Criterion for the prime p to a polynomial g(x) of degree p-1 we need that the coefficient of the term of degree p-1 of g(x) is not divisible by p, the coefficients of all terms of g(x) of degree less than p-1 are divisible by p, and the constant term is divisible by p but not by p^2. The simplest case is when the constant term of g(x) equals p. This is the same thing as saying that g(0) = p. Now we apply this to g(x) = f(x+n). We want p = g(0) = f(n). It is easy to see that f(1)= p. So we try that and it works!

Notably, it is a beautiful illustration of the power of isomorphism, the notion I elaborated intensively in Chapter 2. In other words, the map taking f(x) to f(x+1) is an isomorphism of the polynomial ring $Z[x]$ with itself. Because an isomorphism takes units to units and products to products, it takes irreducibles to irreducibles (R. Smith, personal communication, March 22, 2010).

This could be an example of developing a mathematical argument and reasoning. Students would be well served to keep in mind that as long as we try we will find out either what we expect or something else out of our expectation which tells us either what we should avoid in next try or something new that we may also have interest to give a try.

Based on what we have done in this chapter, we can move a little bit further to **Fermat's Little Theorem** that $n^p \equiv n \pmod{p}$ with p prime and $n \in Z$.

We use **induction**, for example, only for positive n, to start the proof, then from the idea of $a \equiv b \pmod{m} \Leftrightarrow m \mid (a-b)$ we want to show $p|[(k+1)^p - (k+1)]$. So we use **Binomial Theorem** to expand $(k+1)^p - (k+1)$, and then we get $(k+1)^p - (k+1)$ =

$(k^p - k) + p(k^{p-1} + k) + \binom{p}{2} k^{p-2} + \ldots \binom{p}{p-2} k^2$ (*). $p|(k^p - 1)$ as assumed for induction and

then by **lemma 3** just proved above, p divides every term on right hand side of (*), so p

divides the left hand side of (*), i.e., $p|[(k+1)^p-(k+1)]$, as desired. The proof for negative integer *n* is similar with the above one, but the main idea is to prove it for p=2 by induction on *n*, and then do induction for all other prime p, because all prime numbers except for 2 are odd numbers and an odd power of a negative number is still negative, so we can just use the cases for positive numbers.

We may see that the whole proof is directed by simple pure algebraic logic, which I think high school students can master after practice under the guidance of teachers, though Fermat's Little Theorem is a so-called advanced algebra theorem that has never appeared in high school algebra books.

It would be easy to restate Fermat's Little Theorem as that if p is prime, then for any $n \in Z$, $n^{p-1}-1$ is divisible by p. Then this is the basis for the Fermat primality test, which is a probabilistic test to determine if a number is composite or probably prime.

More interestingly, as I noted in Chapter 2.4, Fermat's Little Theorem is at the basis of modern cryptography such as some security coding methods for credit cards, information transfer in banking, ATM machines, electronic commerce, and other secret messages on the internet. The specific algorithm is sophisticated, but here I would like to describe a simple way of encrypting whole numbers.

What can we do with $n^p \equiv n \pmod{p}$ (p prime)? Let $x^b = n$ with $x, b, n \in Z$, and then we get $\left(x^b\right)^p \equiv x^b \pmod{p}$, i.e., $x^{bp} \equiv x^b \pmod{p}$. Then let $a = bp$, so we have $x^a \equiv x^b \pmod{p}$. It seems we need to have another modulo relation between a and b. Working on $a = bp$ seems promising. It is equivalent to $a - b = bp - b$, i.e, $a - b = b(p-1)$. So $a - b \equiv 0 \pmod{p-1}$, i.e., $a \equiv b \pmod{p-1}$. Now we get a slight generalization of

Fermat's Little Theorem, i.e., if p is prime and $a \equiv b \pmod{p-1}$ with $a, b \in Z$, then for any $x \in Z$ we have $x^a \equiv x^b \pmod{p}$. Next we try with specific numbers. Let $p=7$, then $a = 2$ and $b = 8$ satisfy the condition $a \equiv b \pmod{p-1}$. Suppose $x = 3$, then according to the above generalization, we know $3^2 \equiv 3^8 \pmod{7}$. So we have $9 \equiv 6561 \pmod{7}$. Suppose that Amy using our mod-7 algorithm changes the number 3 to 9 by raising 3 to its 2nd power, and then sends to Beth the number 6561 which is equivalent to 9 by mod 7. Beth knows that they are using the mod-7 algorithm and that in order to recover the original number from the number she receives from Amy all she needs to do is to take the 8th root of 6561. Thus, of course, Beth successfully gets the number 3 back. Note that the variant of Fermat's Little Theorem fortuitously helps students see that it is possible that different powers of a same number can still be the same in Z_p (*p* prime), which never happens in Z.

4. Functional Thinking Entailed in Problem Solving and Reasoning

NCTM (2000) urges that school instructional programs should enable all students to—

- Build new mathematical knowledge through problem solving
- Solve problems that arise in mathematics and in other contexts
- Apply and adapt a variety of appropriate strategies to solve problems
- Monitor and reflect on the process of mathematical problem solving (p. 334)

- Make and investigate mathematical conjectures
- Develop and evaluate mathematical arguments and proofs (p. 342)

In this chapter I show how ideas from abstract algebra and high school algebra can be used to engage students in problem solving and reasoning, which entail thoughts about functional relationships. We have seen some examples of isomorphism in the previous chapters, and we will see more in this chapter, yet our focus will be more on the prediction and bijection features of a function. I use examples from permutations (e.g., symmetry group S_3), congruence/isomorphism, dihedral group, certain relation between a group and the cosets of a subgroup, and a little bit concept of group action. In this chapter I still draw on the idea that "[t]eachers should maintain a curricular perspective, considering the potential of a task to help students progress in their cumulative understanding in a particular domain and to make connections among ideas they have studied in the past and those they will encounter in the future" (Martin, 2007, p. 33).

4.1 From labeling a triangle to permutation group

Most high school student may have solved a problem like "How many ways are there to label a triangle using the different letters of the alphabet?" The following is the explanation of the solution of this problem from the point of view of abstract algebra.

If we use A, B, C to label a triangle, we have two distinct ways to do it. ABC means that we start with A on the triangle on the right, go to B counterclockwise, and then go to C and then back to A. This is called a **cycle**. Without moving A, B, and C, BCA means that we start at B, go to C, then to A, and then to B. CAB means that we start at C, go to A, go to B, then back to C again without moving the letters on the triangle. So, all three cycles amount to the same arrangement of letters on the triangle, i.e., ABC, BCA, or CAB equivalently denote the same arrangement of letters on the triangle. We see that C is on our left and A on our right. Similarly, ACB, CBA, or BAC equivalently denotes another same arrangement of letters on the triangle. We see that C is on our right and A on our left. No matter how we cycle the two sets of cycles, the two sets of arrangement of letters on the triangle will never coincide (see Figure 4.1.1). The arrangements ABC, BCA, and CAB actually are in the form of a cycle (ABC) in terms of abstract algebra terminology. The arrangements ACB, CBA, and BAC are also in the form of a cycle (BCA) in terms of abstract algebra terminology.

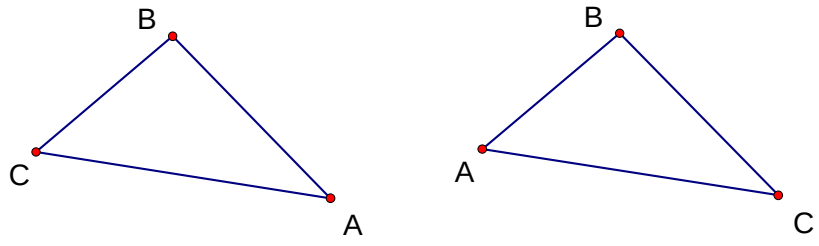


Figure 4.1.1 Two ways of labeling a triangle with three different letters

If we place A, B, and C on a segment, in fact, we will see there are 6 different orderings to arrange the three letters. The linear arrangement ABC means that B is between A and C. Then CBA is the inverse arrangement of ABC on the segment. They are different arrangements of the letters on the segment. Note that on a triangle, B is not between A and C because we can go from A to C without passing through B. Likewise for the two other letters. If we list out all the six distinct orderings of the arrangement of A, B, and C on this segment we get 3×2 different orderings. See the following illustration (Figure 4.1.2).

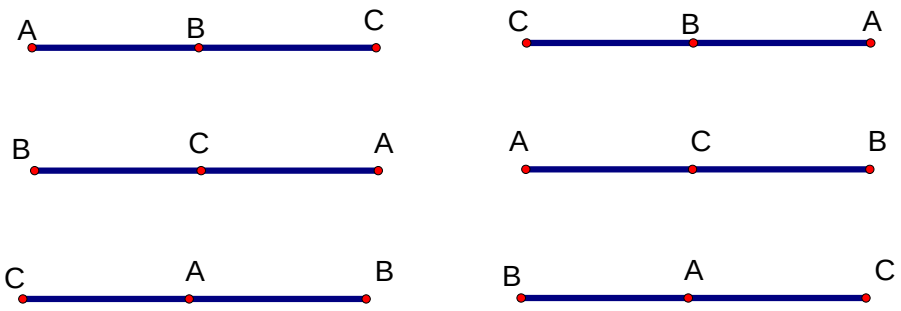


Figure 4.1.2 Six orderings of three different letters on a segment

This "linear arrangement case" actually can be converted mathematically into the symmetries of a group of an equilateral triangle, denoted by T, which is isomorphic to the symmetric group S_3, the set of **permutation**s of three elements, say {1,2,3} or {A,B,C}, or even {Apple, Pear, Orange}, i.e., the set of all *bijections* of a 3-element set with itself. Clearly, the linear arrangement ABC corresponds to the permutation $\begin{pmatrix} A & B & C \\ A & B & C \end{pmatrix}$, or $(A)(B)(C)$ in terms of cycle; the linear arrangement CBA corresponds to the permutation $\begin{pmatrix} A & B & C \\ C & B & A \end{pmatrix}$, or $(AC)(B)$ in terms of cycle; the linear arrangement

BCA corresponds to the permutation $\begin{pmatrix} A & B & C \\ B & C & A \end{pmatrix}$, or (ABC) in terms of cycle; the linear arrangement ACB corresponds to the permutation $\begin{pmatrix} A & B & C \\ A & C & B \end{pmatrix}$, or $(A)(BC)$ in terms of cycle; the linear arrangement CAB corresponds to the permutation $\begin{pmatrix} A & B & C \\ C & A & B \end{pmatrix}$, or (ACB) in terms of cycle. Note that $\begin{pmatrix} A & B & C \\ C & A & B \end{pmatrix}$ namely (ACB) or (CBA) is the inverse element of $\begin{pmatrix} A & B & C \\ B & C & A \end{pmatrix}$ namely (ABC), and the other element is the inverse of itself in group S_3 with its operation of the composition of permutations. The order of S_3, namely $|S_3|$, i.e., the number of elements in the group S_3, is 3!=3×2×1.

Now, how many different triples of letters are there using the alphabet if all three letters are different? Remember, we used only three distinct letters, A, B, and C above. There are 26 choices for one vertex. After that, there are 25 choices for the second vertex. Finally, there are 24 choices for the third vertex. So, there are 26×25×24 ways to choose three letters. But, in this case, the choices are ordered choices which means, for example, that (ABC) is not the same as (CBA) in terms of cycles above. We are making choices as if the letters are in a row on a segment, and as we have discussed ABC, ACB, BAC, BCA, CAB, and CBA are all different because they are arrangements of A, B, and C on a segment or permutations of A, B, and C. So, for each choice of three letters, there are 3×2×1=6 distinct orderings of the three letters as we just got above. Teachers may guide students to see why it is the way to compute the number of elements of a permutation set.

If we choose three letters for the vertices of a triangle, we do not want to order them once we choose them because we have already shown that there are only two of them that are distinct. So, there are (26×25×24)/(3×2×1) distinct choices of three letters of the alphabet disregarding the order. But there indeed are two types of them that are distinct regarding the order, because it is really not an equilateral triangle; (ABC) and its inverse (CBA) are not congruent. So we need to multiply (26×25×24)/(3×2×1) by 2. Hence, there are 2×(26×25×24)/(3×2×1) distinct ways to label the vertices of an irregular triangle with distinct letters of the alphabet.

Before teaching permutations using abstract algebra way I think using traditional numerical methods to approach the solution and to formulize the mathematical thinking elicited from the problem is necessary because it is easier for students to achieve and helpful to understand. These numerical methods can then be extended to the algebraic methods I have described above.

I think the mathematical ideas implied in this problem can be adapted to many other contextualized problems. Teachers may want students in groups to make up some different contexts that fit in the "spirit" of this problem. Hopefully, students will notice there is functional relationship $m = \frac{(n)(n-1)(n-2)...(n-(k-1))}{k!} \cdot 2$ embodied in each context they come up with. In our original context *k* denotes the number of the vertices of the irregular polygon, *n* denotes the total number of choices (e.g., letters of the alphabet) the first vertex can be labeled, and *m* indicates the number of distinct ways to label the vertices of the polygon with distinct letters or notations. If *m* is the dependent variable, *n* and *k* can both be independent variables or students can fix one of them and only leave one to be variable. These setups really depend on the context of the problem students

come up with. “An important convention in algebraic representation is that if there is a predictable relationship between numbers, we generally write one in terms of the other” (Boaler & Humphreys, 2005, p. 31). In short, there are a lot of good ingredients classroom teachers could elicit from this problem to make a worthwhile task of high-level thinking demands.

Note that the relationships cannot be taught explicitly; rather, teachers must use different ways to help each individual student construct and represent the relationships mathematically. Remember that teachers are “modeling for students the practice of asking mathematically worthwhile questions” (Driscoll, 1999, as cited in Boaler & Humphreys, 2005, p. 35). Thus, in some cases, teachers should avoid asking directly, or too early, a question like “What is the relationship between A and B?”. Rather, teachers should help students notice the existence of the relationships themselves. Do not “press the students about this but rather to let them ponder it on their own” (Boaler & Humphreys, 2005, p. 31). Boaler and Humphreys (2005) claimed that the impact of this strategy would surface later in the lesson. It is recommended that “teachers focus students’ attention on what is staying the same and what is changing in order to help them learn how to build rules to represent functions” (Driscoll, 1999, as cited in Boaler & Humphreys, 2005, p. 30). Algebra is a tool for solving and exploring problems, during which manipulation skills can be practiced in an incidental way, but is also a tool for representing mathematical relationships more so than finding results.

4.2 From self-congruence of isosceles triangle to dihedral group

One of Euclid's propositions (Proposition I.5) was that the base angles of an isosceles triangle are equal. Euclid proved it by constructing auxiliary triangles. But here I want to share a different proof of it, which actually is a special case of Euclid's proof. Given isosceles triangle ABC with AB, AC equal, we let A correspond to itself, and let B and C correspond to each other (i.e., B corresponds to C, C corresponds to B). Teachers should encourage students to visualize this relationship as it is an important exercise in visualization. Then we will see AB=AC, $\angle BAC=\angle CAB$, and AC=AB, then $\Delta BAC \cong \Delta CAB$ by (assuming) side-angle-side criterion for congruence of two triangles. Thus, the corresponding angles in each triangle are equal, i.e., $\angle B=\angle C$. We proved the isosceles triangle congruent to itself, with the parts in a different order. It is a fast and kind of tricky proof. Dr. Roy Smith commented that there is a psychological reason that most people find this proof tricky: It involves understanding the concept of a function, not just vertices of a triangle (personal communication, February 9, 2010). In other words, it involves understanding the meaning of the sentence "sending A to A, B to C, and C to B, which in this special case preserves corresponding distances, defines a congruence" as opposed to just roughly thinking one knows what the sentence "these triangles are congruent" means. More specifically, in the Euclidean plane we say geometric figures are congruent if they can be related by a bijective distance-preserving map; at this point it is essentially the same notion of isomorphism.

Yet there is more behind this proof. If we just let A, B, C respectively correspond to itself, it will be the identity self-congruence. If we let the vertices correspond in the way shown above, then it is like a flip self-congruence about the bisector the original angle A.

So there are two self-congruences in an isosceles triangle (only two sides equal). Similarly, an equilateral triangle has six self-congruences: the identity self-congruence, 60-degree and 120-degree rotations about the center of the equilateral triangle, three reflections about the three bisectors of the triangle (Again, teachers should encourage students to picture this in their minds.). Extending this to the square, there are eight self-congruences up to "rigid motions" (translation, rotation, reflection). For a regular pentagon there are 10. Hopefully students would be inclined to ask whether there is any pattern to these numbers and would try to make a mathematical conjecture. As Martin (2007) noted, "[s]tudents must be given opportunities to engage in making conjectures, to share their ideas and understandings, to propose approaches and solutions to problems, and to argue about the validity of particular claims; they must recognize that mathematical reasoning and evidence are the bases for discourse" (p. 46).

There are 2 self-congruences for a triangle with 2 sides equal, and 6=3×2 for a triangle with three sides equal, and 8=4×2 for a regular 4-gon, and 10=5×2 for a regular 5-gon. Thus, there 2n self-congruences (which in abstract algebra textbooks are usually called symmetries) in a regular n-gon. To understand why this is so, we note that a regular polygon with n sides has 2n different symmetries: n rotational symmetries (including the identity symmetry) and n reflection symmetries. In abstract algebra all the symmetries of a regular polygon with respect to the operation of composition of symmetries form the algebraic structure of a finite group, which is called a **dihedral group** denoted D_n. If students could accept that $(Q,\times)$ is a group, then taking all the 2n symmetries composing of a set with the operation of composition of symmetries will help students also accept D_n is a group. However, the difference is that D_n is not a

commutative group while $(Q,\times)$ is. If we first do a 90 degree central rotation to a square and then a central flip, the position of the square should be different than if we reverse the order of the two actions done to it. Incidentally, if we work with the matrix forms of rotations and reflections we will understand those matrices form a non-commutative group. More generally, matrix multiplication is non-commutative and matrices have zero-divisors, so some matrices with the operations addition and multiplication may only form a non-commutative ring.

To some this may look more related to geometry at first sight. But actually this underlines pattern generalization. NCTM (2000) expresses the following expectation on high school algebra that "[i]n grades 9–12 all students should generalize patterns using explicitly defined and recursively defined functions" (p. 296). In this context, the relationship between the number of sides of a regular n-gon and the number of different symmetries in the n-gon (denoted $|D_n|$, the order of the dihedral group, i.e., the number of the elements of the group) yields a representation of the function $|D_n| = 2n$, confining $n \in N$ and increasing from 3. In other words, we need to define the domain and the range for the function. Again, as I emphasized before defining and identifying domain and range of an equation is very important to determine whether the equation is a function. However, most mathematics textbooks, especially at the secondary school level, do not attend to it.

One point I think is worthy of mentioning is that if the aim of the task is to let students find the functional relationship between the number of sides of a regular polygon (denoted by n in the following table) and the number of self-congruences of the corresponding regular polygon (denoted by m in the following table) through figuring out

the growth pattern, it is not suggested to direct students to write down the so-call input-output table as a way to represent the cases they have checked. If we introduce the table like the following one

n	m
3	6
4	8
5	10
6	12

Figure 4.2.1 An example of input-output table

many students will look down the columns for recursive insights because recursive relationships are often stressed in middle school. In other words, they may notice that the first column (n) starting from 3 increases by ones and the second (m) starting from 6 increases by twos. Looking at the recursive relationship within a variable obscures the relationship *between* the variables (independent variable and dependent variable). Students will over-focus on the pattern within each variable rather than the functional relationship between the two variables. This is why I suggested introducing the conjecture in the manner noted above to help students focus on the relevant numbers (the number of sides, the number of rotation symmetries–including the identity symmetry, the number of reflection symmetries) for each regular polygon. By looking at these number patterns, students may be encouraged to visualize the geometric figures. Teachers should give students private think time to ensure that each student has a chance to individually grapple with the task, and then may put students in to several small groups in order to let individual diligence and intelligence contribute to both group and individual success, and later ask students from different groups to draw some different polygons as examples to check their (maybe different) conjectures and share their points of view or explanation

with the whole class. Boaler and Humphreys (2005) noted that students learn a lot when they consider competing ideas, even when some of them are wrong. Furthermore, "verbal statements about functional relationships are critical for understanding … [and] it is important to make sure students verbalize generalizations of patterns before asking them to formalize those generalizations symbolically" (Boaler & Humphreys, 2005, p. 27). So as students try to image the patterns in their mind or draw pictures on the paper and try to verbalize the relationships they get, they are able to understand them more or better.

Moreover, one reason students bog down in the relationship within one variable may be that the set of possible values of one variable (except for the constant function) usually has more than one element or even an infinite number; thus, it is easy for students to move forward along a road when they do not see the end of the road. The relationship *between* two variables is usually not obvious so students must relate two sets of many or even infinitely many elements together to find a unique rule or a unique succession of algebraic operations that tell the interdependent relationship between the two sets. It is as hard as asking students to walk along one road while observing where the other road goes in order to figure out how the two roads are related. In some cases, it might be impossible to use the winding of one road to predict the path of the other one (i.e., there may be no functional relationship between two variables at all). Moving back and forth among the notions of uniqueness, the finite, and the infinite makes it difficult for students to figure out a functional relationship between two sets. And what is also difficult for students to understand and master the concept of function is that it is also hard to say whether a functional relationship exists as an object or as a process; in other words, is a functional relationship something static or dynamic?

4.3 From "Parade Group" to group action

I now introduce the idea of a group action and then relate it to real world experiences that could engage high school students in this mathematical idea.

First I give the definition of **group action** as follows. You may skip this definition, if it looks abstract to you. It won't hurt your reading of the following content. When you need just go back to this.

The **action** of a group G on a set S is given by a homomorphism $\phi : G \to \text{Perm}(S)$, which assigns to every $g \in G$ and every $s \in S$ an element $g \cdot s \in S$, where we normally write $\phi(g)(s) = g \cdot s$, and $\text{Perm}(S)$, as we mentioned before, is a set of bijections from S to itself. Note that since ϕ is a homomorphism we have $\phi(e) = I$, so the operation $\cdot$ satisfies:

(1) $e \cdot s = \phi(e)(s) = I(s) = s$;

(2) $g \cdot (g' \cdot s) = \phi(g)((\phi(g')(s)) = (\phi(g)\phi(g'))(s) = \phi(gg')(s) = (gg') \cdot s$.

Conversely, given (1), (2), the function ϕ defined by $\phi(g)(s) = g \cdot s$ is a homomorphism.

Many projects or initiatives (e.g., Mathematics in Context, Connected Mathematics) advocate curricula with contexts to situate mathematics tasks, which draws students interest, stimulates students' imaginations and investigations, and assists students in making connections among mathematical and everyday concepts that students hold (Middleton & Spanias, 1999). Next I present an example from real life that involves the definition of group and group action. Consider a set of four oral parade commands, which tell the group to pivot and look to a certain direction: stand as you were (S), left face (L), about face (A), and right face (R). The operation is "followed by," which we designate as

F. Interestingly, S, L, A, and R, respectively, correspond to I (the identity rotation of a square), R_1 (the symmetry rotation of a square by 90 degrees), R_2 (the symmetry rotation of a square by 180 degrees), and R_3 (the symmetry rotation of a square by 270 degrees). We can check that the "Parade Set" with the operation F "followed by" really is a group by making an operation table because the closure, associativity, identity, and inverse properties of a group hold for the four commands with the operation "followed by." In fact, the "Parade Group" (denoted by P) is a subgroup of the symmetry group of a square, namely, D_4. So if we notice the corresponding relationships of S, L, A, and R to the elements of the symmetry group of a square, we can immediately realize it is a group. Furthermore, if we consider the set of the front, the left, the behind, and the right of a parade group corresponding to the set (denoted by E) of four edges of a square, say a, b, c, and d, then we can let P act on E. The action is shown in the following operation table:

Action	I	R_1	R_2	R_3
a	a	b	c	d
b	b	c	d	a
c	c	d	a	b
d	d	a	b	c

Figure 4.3.1 "Parade Group" operation table

Thus we can rewrite the elements of P in terms of **permutation cycles**: I=(a), $R_1 = (abcd)$, $R_2 = (ac)(bd)$, $R_3 = (dcba)$. These cycles show mathematically what the elements of the group P are.

Teachers may use this example as a concrete and even a pictorial example to help students get a feel for what a group looks like. Teachers may also show students how a group action can be used to help express its elements in another way. Similarly, we use different words in different contexts to define the same thing. Informally, for example, we say a truck is a vehicle used to carry relatively big and heavy things from one place to another. But if we confine a truck to the context that the Smiths are moving, we may say the truck is a vehicle that carries the Smiths' furniture from their old house to their new home.

Moreover, teachers may use this example to initiate a discussion among students about the relationship between a group and its subgroup. Dweck (1986) noted that appropriately challenging tasks are often the ones that are best for utilizing and increasing one's abilities. I believe students have the abilities to observe, make a conjecture, and then reason it with the help of their peers and teachers, and I want to challenge them to enhance these invaluable abilities.

For example, one of the features of a subgroup is that H (as a subgroup of given group G) is closed under multiplication and its inverse, i.e, $a,b \in H \Rightarrow a \cdot b \in H$ and $a^{-1} \in H$. So if we let any other element, say a_1, of G but not in H multiply any element in H, the element after multiplication will not be in H but still in G. If we multiply a_1 with every element in H, we will get a set a_1H (a **coset** of H), which has no overlap or intersection with H. Then we use another element $a_2 \in$ G neither in H nor a_1H to do multiplication with H. We will get another coset a_2H, which has no intersection with H either for the similar reason as for a_1H. But does a_2H intersect with a_1H? (Some

students may come up with this question.) We claim that if $a_1 \neq a_2$, then applying a_1, a_2 to any two elements h_1, h_2 in H (h_1, h_2 can be the same), respectively, $a_1h_1 \neq a_2h_2$. More generally speaking, two cosets have a nonempty intersection if and if they are identical. What is the reasoning for this claim?

Proof:

"$\Rightarrow$": Suppose b$\in$($a_1H \cap a_2H$). Then there exist $h_1, h_2 \in H$ such that $b = a_1h_1 = a_2h_2$. Thus, $a_1 = (a_2h_2)h_1^{-1} = a_2(h_2h_1^{-1}) \in a_2H$ by closure of a subgroup. So we let $a_1 = a_2h_3$ with $h_3 \in H$, left a_1 multiplying with any other element $h_4 \in H$, and then we get $a_1h_4 = (a_2h_3)h_4$, i.e., $a_1h_4 = a_2(h_3h_4)$. Thus, $h_3h_4 \in H \Rightarrow a_1H \subset a_2H$. Similarly, $a_2 = a_1(h_1h_2^{-1})$, and so $a_2H \subset a_1H$. Therefore, $a_1H = a_2H$, as desired.

"$\Leftarrow$": Common notion: two identical things have nonempty intersection.

Now we see that each element $\alpha \in$ G belongs to only one coset of H, so G can be expressed as the union of distinct cosets αH. If we define the **index** of H in G to be the number of distinct cosets of H in G, denoted [G:H], and define the **order** of H (or G) to be the number of elements of H (or G), denoted |H| (or |G|), then we conclude |G|=[G:H]$\cdot$|H|. Where did this relationship come from? The union of the [G:H] disjoint cosets is all of G. H is in one-to-one and onto correspondence with each coset αH, if we mapping $h \in H$ to αh, so every coset has |H| elements. Therefore, in G there are [G:H]$\cdot$|H| elements all together, which means the same number as |G|.

Based on the example of the "Parade Group", if teachers push students to think a little bit further, they can ask student "What are the relationships you have seen between ｜P｜ (the "Parade Group") and ｜Sq｜ or ｜D_4｜(the square or D_4 group)?" Students

may note that "｜P｜ is smaller than ｜ D_4 ｜." or maybe "｜ D_4 ｜is double of ｜P｜." or "｜P｜ dividing ｜ D_4 ｜equals 2.", or "｜ D_4 ｜is divisible by ｜P｜.". Good. These are all good findings. Then teachers can push students to explore all the subgroups of D_4 by using the definition or properties of a subgroup and ask them to generate the relationship between the order of a group and the order of any of its subgroups. I believe, after going through all the subgroups, students will generate the conclusion that "if H is a subgroup of a finite group G, then ｜H｜ divides ｜G｜". Now it would be the right time to try to discuss whether this "rule" works for any subgroup of any finite group or not (as we just showed above). If students think it is true, then teachers may ask them "Why?", "How do you know?", "Can you prove that?", or "Can you give me some reasons for that?" to put them on the way of "reasoning of justification" (Ball & Bass, 2003, p. 29). Ball and Bass (2003, p. 29) pointed out that "mathematical reasoning is as fundamental to knowing and using mathematics as comprehension of text is to reading". Important learning and mathematical thinking takes place as students try to make sense of something.

To end this section I will elaborate a little more about **group action**. Given a group action G on a set S, and $s \in S$, the **orbit** of s is $O_s = \{s' \in S : s' = g \cdot s$ for some $g \in G\}$; the **stabilizer** of s is $G_s = \{ g \in G : g \cdot s = s \}$. There is a one-to-one and onto correspondence between G/G_s (the set of cosets of G_s in G) and O_s. So $[G : G_s]$ = #(O_s). And then by Lagrange ｜ G ｜ = $[G : G_s]$ ・ ｜ G_s ｜, where G_s is a subgroup of G, we have ｜ G ｜ = # (O_s) ・ ｜ G_s ｜ *, as we suggested students to figure out by themselves in the above example. Once we know this formula *, given the action well

defined, or at least understanding the idea underneath the counting principle, even high school students could learn to count the order of the group G, which may be the simplest question we can ask when we naturally want to know something about the structure of groups we encounter.

Here is a classic example (appreciated and offered by many mathematicians) to illustrate the power of group actions to high school students. It is also a good example to show the conception that groups represent the algebraic version of symmetry (the symmetric group S_3 we talked before in this section was also one fundamental example of a finite group). Let G be the group of rotation symmetries of a cube. How many rotation symmetries are there in G? Let us count as high school students in this way: (1) the identity rotation is 1 element of G; (2) the rotations by 90°, 180°, 270°about the 3 axes through centers of opposite faces count 3×3=9 elements of G; (3) the rotations by 120°, 240° about 4 axes through opposite vertices count 2×4=8 elements of G; (4) the rotations by 180°about 6 axes through midpoints of opposite edges are 1×6=6 elements of G. So there are totally 24 elements in G, i.e., $|G| = 24$. Now let us see how to get this number in a more advanced way by employing group action. There are three useful actions of G we can use: G acts on the set of faces/vertices/edges of the cube. Let us take the action of G on the set of edges of the cube as example. Let S be the set of edges of the cube, so any element of S can go to all the positions of 12 edges of the cube by symmetry rotations. This means the orbit of $s \in S$ is the whole set S, i.e., $O_s = S$ and # (O_s)= # (S)=12. The stabilizer of one edge $s \in S$ is the rotations by 0° or 180° about the axis through the midpoints of the edge and its diagonally opposite edge, i.e., G_s is all the

symmetry rotations that fix the edge $s \in S$, so $|G_s| = 2$. Therefore, $|G| = \#(O_S) \cdot |G_s| = 12 \cdot 2 = 24$. To check this nice result again we can take the action of G on the set of vertices of the cube as follows. There are exactly three rotations leaving a given vertex of the cube fixed, and there are 8 vertices, so there are 8x3 = 24 symmetry rotations in all.

5. Real Numbers and Polynomials

In this chapter I will discuss some topics related to real numbers such as natural numbers, 1-1 correspondence (the bijection feature of an isomorphism), prime numbers, the unique factorization theorem, rational numbers, Archimedean properties, the least upper bound property, Dedekind's axiom, ordered field, proofs of 1.000…=0.999…, the infinite set, the countability and uncountability of numbers, polynomials and its roots, algebraic numbers, and the Cartesian plane over a field. The impetus of writing this section mainly came from reading Dr. Roy Smith's unpublished lecture notes about the construction of the real numbers for the Paideia School at Atlanta, my own class notes for his mathematics classes at the University of Georgia (UGA), and our personal communication around my questions. I gave clear definitions and detailed proofs for most theorems discussed in this section. I believe going through a whole proof is a hands-on mathematical experience and also a good way to reuse and connect the knowledge we have learned and also a process of learning something new. Even though we may not be able to regenerate a whole proof by ourselves without any reference, it would still be easier to pick it up quickly after we get some hints or stimuli, about which most students must agree with me. It is reminiscent of an old Chinese saying "Having it does not mean it is enough, but not having it means it is not enough".

King (1992, p. 37) proposed that abstract algebra is a branch of mathematics that deals with structures having names like 'rings', 'fields', or 'groups' and processing

structures which allow the elements of those sets to be combined in various ways. High school algebra more than likely contains "only rules for the elementary manipulation of number" (p. 37). In other words, high school students view real numbers, for example, just as meaningless numbers operated mechanically, instead of elements of a field, which has its own algebraic structure built out of binary operations on the set itself – again, this is one important idea I accentuate in the thesis. If we ask the question to high school students that in their mind what the differences between real numbers and integers (or natural numbers) are, some of them may blurt out that a real number has infinite decimal digits, but it is not true since an integer also has infinite decimal digits which just are all 0's. Some may say there are infinitely as many real numbers as natural numbers. Unluckily, it is not true either (we will talk about this later). Others may do a better job saying that an irrational numbers cannot be expressed as a ratio of two integers. That is right, but there is in fact more than that to talk about among different number systems to high school students.

Numbers are abstractions intended to capture some quantitative properties of physical objects and to make it possible to compute the answers to various questions; different questions require different kinds of numbers to be useful for answering them; we began historically with the simplest kinds of numbers, positive integers, for the simplest counting problems, and invented new numbers as new problems arose (R. Smith, personal communication, September 9, 2009). For example, the creation of negative numbers was motivated by the demand that the operation of subtraction should work in all cases; the invention of complex numbers was motivated by the question that what is the solution(s) of the equation $x^2 = -1$ when solving for x. Most importantly, the notion

of one-to-one correspondences, which might precede numbers, enriched or developed humans' experience in numbers. Dr. Roy Smith said that he used to use the example of the Cyclops putting aside one rock for each sheep he let out to graze, after Odysseus blinded him, so he would know when they had all come back in the cave (personal communication, February 17, 2010). It was a good example for illustrating the difference between 1-1 correspondences, such as the Cyclops was using, and actual counting, wherein one has a model set of numbers to compare with each other set to be counted. That is, the Cyclops did not know how many sheep he had, but he knew whether the number that went out equaled the number that came back in. Notably, a great contribution by George Cantor (1845-1918) "allowed mathematicians for the first time to come to exact grips with the concept of size of infinities" (King, 1992, p. 185).

5.1 Natural numbers and 1-1 correspondence

When we treat counting numbers as natural numbers, we essentially define **natural numbers** as a set of the number 1 and any other numbers obtained by adding 1 to it repeatedly, assuming the successor of any natural number is different from the number. That is, we define a map f: $n \rightarrow n+1$ from the set of natural numbers N to itself. The **pigeonhole principle** states that, given two natural numbers n and m with $n > m$, if n pigeons are put into m pigeonholes, then at least one pigeonhole must contain more than one pigeon; in other words, m pigeonholes can only hold m pigeons with one pigeon in one hole, adding one more pigeon will require reusing one of the m holes. More mathematically, the principle states that there does not exist an injective function on a finite set whose range (the number of the pigeonholes) is smaller than its domain (the

number of the pigeons); in other words, there cannot exist an injective function from a larger finite set to a smaller finite set. Hence, the pigeonhole principle implies that if S is a finite set, then every injection f: S→S is a surjection. Thus, its (equivalent) contrapositive says that if there is an injection f: S→S which is not a surjection, then S is not finite, namely, infinite. Therefore, the map f: $n \rightarrow n+1$ from the set of natural numbers N to itself as we defined above actually is an injection but not a surjection since $n \neq 0 \Rightarrow n+1 \neq 1$. Hence, N is an infinite set, the existence of which we assume here. Moreover, if we consider any infinite set S, we can similarly find an injection f: S→S which is not a surjection. So we may claim a definition of an infinite set that: a set S is infinite if and only if there is an injection f: S→S which is not a surjection. In other words, a set S is infinite if and only if there is a bijection (one-to-one correspondence) between S and a proper set of S. This, in fact, is the definition of a **Dedekind-infinite** set. The usual definition of an infinite set is that: a set *A* is infinite if it cannot be put in bijection with a finite ordinal, namely a set of the form $\{0,1,2,...,n-1\}$ for some natural number *n*. So an infinite set is one that is literally "not finite", in the sense of bijection. Most modern mathematicians assumed that a set is infinite if and only if it is Dedekind-infinite.

If it is possible to figure out some way to order a set of things, such as defining natural numbers, there is a first one, then a second one, then a third one, and so on, then we really can put our whole collection of things down in this list so that they all get counted. If this can be done we call the set **countable**, even though it may be infinite. So the real problem for counting infinite collections is to find the right way to order or enumerate them. Here is a general principle for recognizing countable infinite sets: Suppose that S is a collection of things which can be broken up into a countable number

of sub-collections, S1, S2, S3, S4, ..., and so on, each having a finite number of elements, and then S is a countable set. For instance, rational numbers (i.e., fractions) are countable. Teachers may want to engage students in finding a way to order or list all the rational numbers (which has been described in a lot of books). Here is the brief idea for a complete list of all positive rational numbers: {1/1, 1/2, 2/1, 1/3, 3/1, 2/2, 1/4, 4/1, 2/3, 3/2, 1/5, 5/1, 1/6, 6/1, 2/5, 5/2, 3/4, 4/3, 1/7, 7/1, 3/5, 5/3, 1/8, 8/1, 2/7, 7/2, 4/5, 5/4, ...} (observing the sum of the numerator and the denominator of each fraction).

Now, based on the notion of countability, we claim that the cardinality of the set of all even (or odd) numbers equals that of the natural numbers, by the definition of "same cardinality", namely that there exists a bijection between the sets. So it is not as many students think that the size of the set of natural numbers are twice big than all (infinitely many—easy to prove) even (or odd) numbers, because there exists a bijection $n \rightarrow 2n$ (or $n \rightarrow 2n-1$) where n is any natural number. What about the set of prime numbers? Are there infinitely many prime numbers? Nobody has found a map to get all the possible primes from natural numbers, but mathematicians did come up a classic way to prove the **fact** that there are infinitely many prime numbers. Once we know this and the **theorem** that every infinite subset of natural numbers has a 1-1 correspondence with all natural numbers, we will see that there are as many primes as natural numbers, since primes form an infinite subset of natural numbers. Before we prove the fact and the theorem we just now stated above, we need to introduce some fundamental and important mathematical ideas.

The **well-ordering principle**: Every non-empty subset of positive integers (*N*) contains a smallest element. Though students would intuitively agree with this statement, the mathematical ideas entailed in the following two proofs need students' attention.

Proof.

Let S be a non-empty set of positive integers. If $1 \in S$, then clearly $1 \in s$ the smallest element. If not, let T={t$\in$ *N*: none of the positive integers 1, 2, 3, … , t belongs to S}. Suppose k$\in$ T. If k+1$\notin$ T, then k+1 is the smallest element of S. If, however, for all k$\in$ T it is true that k+1$\in$ T, then by the principle of mathematical induction, T=*N*, and so S=*N*-T=ϕ, i.e., S is empty, contradicting the hypothesis.

Proof by contraposition.

That is to prove a subset of the positive integers with no least element is empty. Let S be a subset of the positive integers *N* with no least element. Define T = {the complement of S} = *N*-S. Claim: T = *N*, i.e., S=ϕ.

By strong principle of induction, i.e., a subset T of *N* equals *N* if it satisfies the two properties (i) 1 is in T, (ii) if {1,...,k} is in T, then k+1 is in T.

Ok, certainly 1 is in T, since if not, and 1 is in S, then S would have a least element, namely 1, contradicting to the hypothesis.

Now suppose {1,...,k} is in T, and ask whether k+1 is in T. Similarly as the above step, if not, then k+1 is in S, and would also be the least element of S, since none of the smaller elements {1,....,k} lies in S, so it is again a contradiction to the hypothesis.

Hence by strong induction, T = *N*, so S=*N*-T=ϕ, as claimed.

Typically each proof uses **induction**, a basic technique in mathematics. The first proof does not offer as many details or as much clarity as the second one, so it might be a

little harder to follow the first one without having gone through the second one when reading them in the order I have presented them. Hence, it is important to keep in mind that the way a proof is presented to students affects their understanding. Though the two proofs essentially are the same, the first one proves the original statement by discussing successive possible cases one by one while ending the last case by contradiction; the second one instead proves the contrapositive of the original statement and uses contradictions to justify the satisfaction of the two properties of induction. Note that this fortuitously reflects the equivalence between the ideas of proof by **contradiction** and **contraposition**. Incidentally, this principle does not hold for the positive real numbers. For instance, the infinite set of real numbers of form {1/2, 1/3, 1/4, 1/5, ... } contains no least element and its greatest lower bound is zero, but that is not in the set.

The **Unique Factorization Theorem of natural numbers** (namely, **Fundamental Theorem of Arithmetic**): Every positive integer greater than one can be factored into positive prime factors in exactly one way.

(The usual proof is by induction. Here I prefer the narrative proof which I got from Dr. Roy Smith (personal communication, September 6, 2009) because I think it does a better job of expounding on the mathematical thought.)

Proof.

(1) The existence is the easy part.

Suppose there is an integer x>1 that does not factor into primes, then by the **well-ordering principle**, there is a smallest such positive integer, i.e., there is an x>1 that does not itself factor, but such that every smaller integer greater than 1 does factor into primes. But this is impossible. Our number x cannot itself be prime (with just one

prime factor), or else it would already be factored into primes, so it can be factored into two factors both of which are smaller than x and greater than 1. But then both of the factors, being smaller than x, will have prime factorizations, by our assumption, and then by putting the two factorizations together, we would get a prime factorization of x. This contradiction shows that no such smallest non-factorable number x exists, and hence indeed no non-factorable numbers exist at all.

(2) The uniqueness part is harder, and may have been proved first by Gauss.

The **key lemma** usually used nowadays, is to show: if a prime integer p divides a product of positive integers ab, then p divides either a or b.

A nice way to do this follows from a classic fact about **greatest common divisors (g.c.d.)**. It can be stated as a fact about measuring lengths using two different rulers, which are commensurable, i.e., whose ratio of lengths is a rational number. The basic result is that the shortest length one can measure by using both rulers, equals the longest length that can itself be used to measure both rulers. Algebraically, given two integers a,b, the smallest positive integer that can be written in the form na+mb, where n, m are any integers, either positive or negative, equals the largest integer d such that d divides both a and b evenly.

Assuming this, if p is prime and divides ab but does not divide a, then the largest integer that divides both p and a is 1, hence 1 can be written in the form of 1=na+mp for some integers n, m. Then multiplying by b gives us b=nab+mpb.

Now assuming p divides the product ab, it follows by **3-term principle** that p divides both terms on the right side of the equation, hence divides also the left side.

(Incidentally, this is another good illustration of the equivalence relation indicated by

the equal sign "=". The previous example is in Chapter 3) Thus we have shown the above **key lemma** that if p divides ab but does not divide a, then p divides b.

Now suppose we have factored some number n into primes in two ways $n = p_1 p_2 p_3 \ldots p_r = q_1 q_2 q_3 \ldots q_s$ where r and s are some positive integers. Since p_1 is prime and divides the left side it also divides the right, hence by the **key lemma** p_1 divides some q_i, which we may renumber as q_1. But q_1 is prime so if p_1 divides it, since $p_1 \neq 1$, it must be that $p_1 = q_1$. Then we can cancel p_1 and q_1 on both sides and have a new equation $n = p_2 p_3 \ldots p_r = q_2 q_3 \ldots q_s$. We now apply the same argument to the prime p_2, eventually canceling it with some prime q_i we may renumber as p_2. Eventually, we have canceled all primes on both sides, in particular, each prime p_j was equal to some prime q_k, and so the factorization was unique.

Thus, the above two parts constitute the whole proof.

Now let us prove the aforementioned **fact** that **there are infinitely many prime numbers**.

Proof by contradiction.

Suppose there are finitely many prime numbers, which are $p_1, p_2, p_3, \ldots p_m$, where m is a positive integer. Now consider the number $M = p_1 p_2 p_3 \ldots p_m + 1$. Since M is not any of the already known primes, by the **Unique Factorization Theorem / Fundamental Theorem of Arithmetic**, it is can be factored as a unique product of primes. (In fact, this needs only the existence part of the theorem, i.e., only for this proof whether the factorization of

M is unique does not matter.) But none of the already known primes divides M. We reach a contradiction, so there are infinitely many prime numbers. Done.

By the way, Euclid's definition of primes is "A prime number is that which is measured by a unit alone" (Definition VII.11) and Euclid proved "Prime numbers are more than any assigned multitude of prime numbers" (Proposition IX.20) by working on a special case when m=3. In addition, the public lecture Lang (1985b) gave about prime numbers on May 16, 1981 is really a great illustration about the interactions among primes, integers, logarithmation and exponentiation, limit and infinity, discreteness and continuity. I recommend it to those who have interest in further reading.

Next we prove the aforementioned really important **theorem** that **every infinite subset of natural numbers has a 1-1 correspondence with all natural numbers**.

Proof.

Let S be any infinite subset of natural numbers N. By the **well-ordering principle** S contains a smallest element, denoted s_1. Let s_1 correspond to 1. Then we take s_1 off from S, so again the new set $S-\{s_1\}$ has a smallest element, denoted s_2. Let s_2 correspond to 2. Then we take s_2 off from $S-\{s_1\}$, so the set $S-\{s_1, s_2\}$ contains a smallest element, denoted s_3. We now apply the same argument to s_3. Repeating this process, we eventually can make a 1-1 correspondence between every element of any infinite subset of N and every element of N.

Incidentally, how do we know all the natural numbers we defined exist? That is, how do we know the set of positive integers exists as a whole? It would be an infinite collection, and there is, in our physical world, no infinite collection of anything.

According to Gamow (1957, pp. 5-6), in order to show the existence of numbers which exceed not only the number of grains of sand which would even equal to a mass the size of the universe, Archimedes (in his treatise *Sand Reckoner*) set out to determine an upper bound for the number of grains of sand that fit into the universe by inventing a way to talk about extremely large numbers, which is similar to the way large numbers are written in modem science, and estimating the size of the universe according to the then-current model. If we think about it, not only the grains of sand on all the beaches and oceans are finite in number but even the total number of atoms in the universe is finite. Though we could not directly observe anything infinitely many (few) or large (small), "all our observation does take place in space and time and is of what is spatially and/or temporally extended" (Tiles, 1989, p. 21). Hence, we need a way to think coherently about infinity and our observations in space and time (continuous magnitudes). In our mind we believe that there are some really infinite numbers, which are larger than any number we can possibly write out no matter how long we work. "Thus 'the number of all numbers' is clearly infinite" (Gamow, 1957, p. 14). At least we have arrived, (some of us anyway), at an act of faith by which we assume the existence of an infinite collection of things called the counting numbers or natural numbers, which we assume to have the familiar arithmetic properties we know to hold for the few relatively small numbers we actually have used in our lives (R. Smith, personal communication, September 6, 2009).

5.2 Constructive and axiomatic approaches for Real numbers

In this section I will discuss the construction of the real numbers first from a constructive approach (inspired mostly by Dr. Roy Smith's Paideia lecture notes) and

then from an axiomatic approach. In addition, I will share another constructive approach (Cauchy Model) offered by Dr. Roy Smith in an email in the appendix.

A closely related concept to that of a number is the concept of a "numeral", which is a symbol that is used to represent or to "name" a number; a number is more of an idea, whereas a numeral is more of a physical object, something we write on paper for instance (R. Smith, personal communication, September, 6, 2009). We sort of have names for all of the positive integers, so as the negative integers (just changing the direction of counting), at least in the sense that we know how to proceed along from one to the next. It is not harder to represent the rational numbers, those useful numbers that provide solutions to equations like $bx-a = 0$, where a, b are integers, and $b\neq0$. Thus, we just define a rational number to be given by a pair of integers (a, b) where $b\neq0$, but we agree that the two pairs (a, b) and (c, d) shall represent the same rational number if (and only if) $ad-bc=0$. We also write a/b, as a fraction, of course, for the number represented by (a, b). So to name a rational number we just need two integers and thus a finite number of the basic digits. Also the difference between a number and a numeral is pointed up by the fact that 2/3 and 4/6, for example, are two different numerals (or names) for the same rational number. However, one of the difficulties in conceiving of the set of all real numbers is the problem of finding names or numerals for all of them. There are just too many real numbers for us to be able to easily give names to all of them.

5.2.1 The constructive approach

We have already met the phenomenon of using (infinite) decimal expansion to represent a real number. Let us assume it here to start our **constructive approach** for the

real numbers. Then I suggest students think about the following question: “Can you explain why a real number is rational if and only if its decimal representation is either finite or repeating?” (c.f., Shifrin, 1996, p. 56, Exercise 11). As we know the rational numbers are defined as the “fractions” formed by taking “ratios” of integers. Apparently the decimal representation of a fraction is either finite or repeating. If a real number r has finite decimal representation, for example, r has n decimal digits, then r can be expressed in form of $\frac{r.10^n}{10^n}$ which is a fraction since $r \cdot 10^n$ is an integer. What if a real number r has repeating decimal representation, for example, the decimal representation of r is in form of m random digits followed by n repeating digits? Let us put this into a concrete example illustration as follows. Take $r = 0.432\overline{5678}$. Thus, $r = 0.432 + 0.000\overline{5678}$

$= 0.432 + 0.0005678 + 0.00000005678 + 0.000000000005678 + \ldots$.

Every item in the above expansion is a finite decimal, though there are infinitely many finite decimals of such kind. We have just now shown any finite decimal is a rational number. So the sum of rational numbers (infinitely many finite decimals) will still be a rational number, since Q, as a ring or a field, is closed under addition. In high school, students usually are told what sort of numbers are called rational numbers, and then those real numbers not belonging to Q are called irrational numbers. But actually what are real numbers? Students are seldom challenged to reason about why the definition is what it is.

If we treat real numbers as one-to-one correspondence points on the real line, as taught in high school, there would be a historical perspective to approach this concept. The "real" numbers are designed to solve the problem of measuring lengths of line segments; there is nothing any more “real” about them than there is about any other kind

of numbers; all numbers are imaginary constructs; the difference between rational numbers and irrational numbers is much greater and subtler than the difference between real numbers and imaginary numbers (R. Smith, personal communication, Fall 2009). Measuring lengths is not such a simple problem and consequently the real numbers are rather sophisticated things (later in this section 0.999…=1 is a specific example of this). We choose a point x, and assume for simplicity that it lies to the right of zero. Then lay off copies of the unit interval end to end, on the line, starting from zero, until we get one, whose right end point does not go to the right of x, but so that the right end of the next unit interval does go to the right of x. The number of unit intervals we have laid off, i.e., the largest number that do not reach actually to the right of x, is the integer part of the decimal we are constructing. Then we get the tenths part in a similar way; i.e., subdivide the unit length into ten equal parts, and then take one of these tenths and start laying off copies of it end to end starting from the point marking the integer part of x. Again, there will be a segment which does not itself reach to the right of x but such that the next segment will do so. The greatest number of segments that do not reach to the right of x is a number between zero and nine, called the tenths digit in the decimal expansion of x. We get the hundredths digit the same way, and continuing, we can construct as much of the infinite decimal as we want. Now, of course, we think it is obvious that this procedure can be carried out, but in fact in doing so we have tacitly endowed the real line with a special property, the **Archimedean Property I**:

Given any point x on the real line, and any positive length y (determined as the interval between two distinct points), if we lay off enough copies of the given length, starting from zero on the line, we will eventually (in a finite number of steps) go past x.

Algebraically, given any two positive real numbers x, y, there is an integer n such that ny>x.

By our geometric statement this property is apparently intuitively true in Euclidean geometry.

Algebraically, if x is any real number then there is an integer n which is bigger than x. (c.f., Hartshorne, 2000, p. 115 & p. 139)

Proof. Just take n to be one more than the integer part of x. Or, take any positive length y equal 1 the unit length. Done.

The **Archimedean Property I** says that x cannot be "infinitely far away from zero". But what about the possibility that x is “infinitely close to zero”? So we claim the **Archimedean Property II**:

If x is any point to the right of zero, and if we consider any other finite interval extending to the right from zero, such as the unit interval, then it is possible to subdivide that interval so that the first subdivision will occur between zero and x. Algebraically, if x is any real number bigger than zero, there is a positive integer n such that 1/n is smaller than x.

Proof.

To see this geometrically, assume first that we are dealing with the unit interval, just so the interval will have a name. Then consider the case where the end of the unit interval, the unit point, is already to the left of x. Then there is nothing to do, and we are finished. So assume now that the unit point, call it 1, is to the right of x. Now assume the **Archimedean Property I** to find n so large that n copies of the interval from 0 to x, whose right endpoint we will call nx, reaches past 1. Now just subdivide both intervals [0,

nx] and [0, 1] into n equal pieces. Since nx is to the right of 1, the first subdivision of [0, nx], which is x, should be to the right of the first subdivision of [0, 1], which is what we wanted to prove. Dr. Roy Smith pointed out that the word "should" is in there because we do not see immediately how to prove that from Euclidean geometry, so either work it out ourselves or assume Archimedean property II as an axiom if we want to (personal communication, September 6, 2009).

Of course, to prove this algebraically we just take n greater than 1/x. If the integer part of x is bigger than 0, then 0.1=1/10 (where n=10) is smaller than x. That takes care of that case. If the integer part of x is zero, just go out until we find a non-zero digit (say, *k*-th decimal digit) in the expansion of x. Then the number, which has all zeros one place farther out than x does, but then has a 1, and all zeros after that, is smaller than x and has the form $1/10^{k+1} = 1/n$, where $n = 10^{k+1}$. Done.

Corollary II.1: If x is any positive real number (point to the right of zero), then there is a positive integer m such that $1/10^m$ is smaller than x.

Proof.

Since $1/10^m < 1/m$ for any positive integer m, but as shown in the above proof $1/m < x$, then $0 < 1/10^m < x$. Done.

Corollary II.2: If x is a non-negative number which is smaller than 1/n, for every positive integer n, then x is zero.

Proof.

This is a rephrasing of the **Archimedean Property II**, which says if it were not zero then some number of form 1/n would be smaller than it. But by the hypothesis, x has to be 0.

Even though we cannot construct the entire infinite decimal in a finite amount of time in the way we showed before, neither can we actually construct the whole of the familiar collection of positive integers in that we have to just put dots eventually when we were trying to write down all of them, but we know how to construct them as we wish. In this situation, most mathematicians are content to say that we have constructed, or at least that we have given a prescription for constructing, the full decimal expansion of x, and to leave matters there (R. Smith, personal communication, September 6, 2009).

Thus, we know how to assign the real number (an infinite numeral representing) to a point on the real line, with which people extend their conceptions or considerations of "countable" things to "measurable" things. This reminds me of an interesting line from an ancient Chinese poem written by Han, Yu (an essayist and poet during the Tang Dynasty), which I translate into "I am crooning verse alone; there are extremely little pieces of sadness in my heart, which no one can trim for me." Han was using a metaphor that assigns a magnitude (let us say a length on a number line) to each of his very little pieces of sadness. Though the length is extremely small, he still doubted that people can reduce or shorten it. Let us consider it more carefully in certain mathematical system. Suppose we can sign an infinite decimal to the length of Han's sadness which satisfies the Archimedean Property (A) like the real numbers and the length is very, very small which approaches 0. With this hypothesis, we are sure we can trim down his sadness for him, i.e., it is true unless (A) does not hold for is sadness.

Now, what about the reverse? That is, suppose we start from an arbitrary infinite decimal, is it necessarily the real number coming from some point on the line? The answer is essentially yes, with one class of exceptions: certain points on the real line, the

ones having finite decimal expansions, also have another different expansion which is infinite. For example, a number like 19.85000000000..... , has another expansion namely 19.8499999999..... . However, the method that we described of assigning a decimal to a point on the line will not allow us to come up with the second of these two expansions. Thus we could say that, by our construction, the points of the real line correspond exactly to those infinite decimals that never become eventually equal to all nines. It will be better though, if we go ahead and think of real numbers as given by all infinite decimals and just remember that every infinite decimal that ends in all zeros (i.e., every "finite" decimal, or "terminating" decimal) can be written again, in a different way, as an infinite decimal that ends in all nines. This points up the fact that the decimals are numerals, i.e., merely names for the numbers and not the actual numbers themselves; some real numbers have more than one name -- that is it (R. Smith, personal communication, September 6, 2009).

One thing which I think really requires the attention of both students and teachers is that why 0.999… and 1.000… are two different names for the same real number 1. Calculator is convenient but affects students' understanding of real numbers. Calculator rounds the real number 0.999… to a rational number, say, 0.999999. In this case, the calculator only does arithmetic computation of finite decimals, so students will never get the result (0.999…)(0.999…) = 1. No wonder it is hard for students to believe mathematically 1.000… = 0.999…, while they believe 1/3 = 0.333… or 1/9 = 0.111… (*), and nearly no high school students realize that multiplying (*) on both sides by 3 or 9 yields 1.000… = 0.999… . And note that students usually think $m.a_1a_2a_3... < m+1$ where m is an integer, but it turns out to be not right when all $a_i = 9$ with $i \in N$.

We have seen one proof of 1.000… = 0.999… before in Chapter 3. Now I will give a proof of 1.000… = 0.999… by using the **Archimedean Property II** or its corollary.

Proof by contradiction.

If not, then apparently we suppose 1 > 0.999…. And so 1-0.999… = X which is greater than 0. Hence, by **Corollary II.1** above, there exist a positive integer *n* such that $0 < 1/10^n < \mathrm{X}$. Thus, we have $\left(0.999... + \frac{1}{10^n}\right) < (0.999... + \mathrm{X})$, i.e., $\left(0.999... + \frac{1}{10^n}\right) < 1$.

But the truth is that for every integer $n>0$ $\left(0.999... + \frac{1}{10^n}\right) > 1$. For example, when $n = 2$ $\left(0.999... + \frac{1}{10^2}\right) = 0.999... + 0.01 = 1.00999...$, which is greater than 1. So we reached a contradiction. Therefore, 1.000… = 0.999… is true! Done.

Teachers may want to use this little nice proof to promote students' appreciation of the value of logic. Well, there is another proof by doing some basic algebraic operations, which high school students may like better: Let x = 0.999999… . Then 10x = 9.999999… . So 10x – x = 9, i.e., 9x = 9, on both sides of which cancelling 9 (non-zero) yields x = 1.

For later use, we define a set S in a vector space over R is a **convex set** if and only if whenever x and y in S the entire line segment $\overline{\mathrm{xy}}$ lies in S. A convex subset of a line is sometimes called an **interval**.

In order to prove the upcoming "Squeeze Theorem" and its corollary, let us prove a stronger version of the Archimedean axioms, namely,

the **Archimedean Property III (Density of rational numbers)**:

Between any two distinct real numbers there is a rational number. In other words, if x and y are positive infinite decimal numbers and $x<y$, then there is a finite decimal a that lies between them in the sense that it is greater than x and less than y, i.e., such that $x < r < y$. (This says x and y, if different, cannot be infinitely close to each other, as measured by our unit interval.)

Proof.

First make sure neither x nor y ends in all nines. Then, suppose x and y agree out through the n^{th} decimal place but y is larger in the $(n+1)^{st}$ place (trying out one example helps students understand). Since we may assume x does not end in all nines, it is possible to go out further than the $(n+1)^{st}$ place and find an entry in x which is less than a 9. Let r be the finite decimal obtained by letting r agree with x out to that digit but then replace that digit by the next larger digit, by the **Archimedean Property I**. Then we complete r by putting all zeros after that. Then r is greater than x but less than y. Done.

(I used a pretty fundamental and intuitive method to prove this axiom. Some may want to refer to an algebraic proof given by Shifrin (1996, pp. 53-54).)

Note that this may be an explanation to students' questions like "What is the previous (or next) number before (or after) 1/2?" noticing the answer also depends on which number systems are talked about.

Interestingly, please imagine that we are walking straightly along a line, and there exists a gap on the line, namely, an interval [x, y], whose length is 1/m (m is any given natural number), wherever in front of us. That is, if where we are standing now is the origin and facing the positive *x*-axis, then $0<x<y$. As long as the length z of our constant walk step is less than 1/m, we will eventually fall in that gap. We can take z=1/n (natural

number n greater than m), then we know there exist a number of steps we walk by, say k steps, such that $x<(k/n)<y$. Similarly, if we are walking in a limited room, if there is a round hole with diameter of length 1/m (as defined above) and we walk in constant step of length z (as defined above), and if we keep walking in the room, then we will, again, eventually fall in that hole.

Squeeze Theorem: If x and y are any two real numbers, and if A1, A2, A3, ..., and B1, B2, B3, ... , are two infinite sequences of finite decimals such that we have $An\leq x\leq y\leq Bn$,for all natural numbers n , and such that (Bn-An) approaches zero, as n approaches infinity, then x=y.

(Note that teachers may guide students to define a sequence approaching to zero by using algebraic language.

Definition: A sequence {Sn} $n\geq 1$, of real numbers, is said to approach zero, as n approaches infinity, if and only if, for any given positive real number ∂ there exists a corresponding positive integer N, such that for every $m\geq N$, we have $|Sm|\leq \partial$. In shorthand, we write $Sn\rightarrow 0$, as $n\rightarrow$infinity.)

Proof by contradiction.

If not, so that in fact $x < y$, then the **Archimedean Property III (Density of rational numbers)** would give us a finite decimal c such that $x < c < y$. Applying the **Archimedean Property III** again gives us another finite decimal d such that $c < d < y$. Then we would have $An < c < d < Bn$ for all n, where now all these numbers are finite decimals. Subtracting c from the last three of them now gives $0 < d-c < Bn-c$ (1). On the other hand, subtracting An from the first two of them gives $0 < c-An$, and adding Bn-c to this gives $Bn-c < (Bn-c)+(c-An)$, i.e., $Bn-c < Bn-An$ (2). Putting (1) (2) together gives $0 <$

d-c < Bn-c < Bn-An, for all n. This, however gives a contradiction, since then ∂ =(d-c) would be a positive number which (Bn-An) always stays greater than, in contradiction to the assumption that (Bn-An) "approaches (arbitrarily near to) zero".

This **Squeeze Theorem** directly yields the following **Corollary**:

If [A1, B1], [A2, B2], [A3, B3], ... is an infinite sequence of closed bounded nested intervals, whose end points are finite decimals, and whose lengths, (Bn-An), approach zero, then there is *at most* one real number which is in all of the intervals.

Okay, let us now start from any infinite decimal at all, possibly even one that ends in all nines, and explain how to construct the corresponding point on the real line. We just give an example: say the decimal is given by 16.16116111611116111116... (adding an extra one each time before the next six). This infinite decimal should really be thought as an infinite sequence of finite decimals: 16, 16.1, 16.16, 16.161, 16.1611, 16.16112, 16.161161, ... , and so on. That is, the **sequence** of those rational numbers **converge**s to that real number, which in calculus is the **limit** of the sequence**,** or the **least upper bound** (which we will define in a little while) of the set constituted by all the decimals in the sequence. That real number actually as a "**cut**" of the real line divides the rational numbers into two classes (We will talk more about this soon later). We may image that the "guy" 16.16116111611116111116... (a real number in terms of decimal expansion) has an infinite long name, with 16 as its 1st name, 16.1 its 2nd name, 16.16 its 3rd name, 16.161 its 4th name, ... , and so forth. To this sequence of finite decimals we associate the following infinite sequence of closed bounded intervals on the real line for which these decimals are the left end points: [16, 17], [16.1, 16.2], [16.16, 16.17], [16.161, 16.162], [16.1611, 16.1612],....., and so on. Note that the length of succeeding intervals is

growing shorter: the 1st interval has length $1 = 1/10^0$, the 2nd has length $0.1 = 1/10 = 1/10^1$, the 3rd has length $0.01 = 1/100 = 1/10^2$, and so on. Note also that this is a "nested" sequence of intervals, in the sense that each interval completely contains the next interval in the list. That is, any point that lies in any one of these intervals automatically lies also in every previous interval in the list. Then we assert that there is exactly one point x on the line, that lies in every one of these intervals, and that x is the point corresponding to the infinite decimal number we started with. This assertion indeed is the **Cauchy completeness property of the real numbers**:

If [A1, B1], [A2, B2], [A3, B3], ... is any infinite sequence of closed, bounded, nested intervals, whose endpoints are finite decimals, and such that the sequence of their lengths, (Bn-An) with n=1, 2, 3, … , approaches zero, then there is exactly one real number which lies in all the intervals.

Note that if those nested intervals are unbounded or unclosed, by the **Archimedean properties** when n approaches infinity there is no common point in all the intervals.

In the above discussion, we have actually used the conception of **ordering** of real numbers. In other words, if all digits of two numbers (decimals) are correspondingly equal (except for those ending in all 9's and all 0's, i.e., modulo a certain equivalence relation such as 0.999…=1.000…, about which we have talked before), then the two decimals are equal. Starting from the first digit of one number and stopping at the first digit that is different from the corresponding one of the other number, the number having bigger first different digit is bigger.

Then we come to the question about how to **add** two positive real numbers if their decimals are infinitely unrepeated. I believe all high school students know how to add

two rational numbers. So then we can use the aforementioned idea that an infinite decimal can be treated as a convergent infinite sequence of nested closed bounded intervals with rational (finite decimal) endpoints on the real line. Therefore, the sum of the two infinite decimals will be represented by the sum of the two convergent infinite sequences of intervals (as for the precise choice of the sequences, please refer to the Cauchy Model stated in the appendix). Similarly, we can check that the new sequence converges to exactly one common point that is the sum of the two real numbers. Based on this idea we can construct the summation, multiplication, and division of any two real numbers and then check those basic algebraic laws because rational numbers form a field.

Thus we may say that the real numbers in terms of decimal expansions with respect to the operations "addition" and "multiplication" form an *ordered* field.

As we have checked that our infinite decimal construction of real numbers satisfies Archimedean properties, next we check that infinite decimals have the **least upper bound property (LUB)**: Every nonempty set S of real numbers, which has an upper bound, has a least upper bound.

Before giving the proof let us look at the relevant definitions as follows.

Upper bound: Let S be a set of real numbers, and call b an upper bound for S if b is a real number and if b is at least as big as every number belonging to S. (Particularly, if S is the empty set then every number is an upper bound for S.) **Least upper bound (l.u.b.)**: If S is a set of real numbers, a least upper bound for S is a number L such that L is an upper bound for S, and no number smaller than L is an upper bound. (Note that it is not required that the l.u.b. should itself belong to the set S, but only that it shall be a real number.)

Proof of LUB.

Here, we may consider only the case when our numbers are all positive (easier for students to think about). If some of our numbers are positive and others are negative then we only need to consider the positive ones, since positive is greater than negative. If our numbers are all negative, we apply the similar ideas to all positive numbers to argue.

We think of the real numbers as given by infinite decimals, as usual. Note first that since our set S is bounded above, some integer is an upper bound. Moreover since the set is nonempty, some integer fails to be an upper bound. Thus there is a smallest integer which is an upper bound. This is not of course necessarily the (real) least upper bound we are looking for, since it may not be the least real number which is an upper bound. Now we define the least (real) upper bound x of the set S one digit at a time. To define the integer part, take the largest integer that is not an upper bound for S, i.e., one less than the smallest integer upper bound. Now among the numbers {n.0, n.1, n.2, n.3, n.4, ..., n.9} where n is an positive integer, there is exactly one such that it is itself not an upper bound for S, but such that we get an upper bound by adding 0.1 to it. Let a_1 be the digit between 0 and 9 in which this number ends. Then we state that our number x starts out $n.a_1$. Now among the numbers $\{n.a_10, n.a_11, n.a_12, \ldots, n.a_19\}$ there is again exactly one such that, it is not itself an upper bound for S, but such that we get an upper bound by adding 0.01 to it. Let a_2 be the appropriate digit, and then our number x starts out as $n.a_1a_2$. Continuing in this way we get at least a prescription for constructing an infinite decimal $x = n.a_1a_2a_3a_4\cdots$, and one which I claim is the least (real) upper bound of S.

Hence we must check two things:

(1) x is an upper bound for S

(2) No number smaller than x is an upper bound for S.

For (1), Suppose x is not an upper bound for S, then there would exist a number z in S which is larger than x, i.e., x<z and z∈ S. Now if $x = n.a_1a_2a_3a_4\ldots$, then write $x_m = n.a_1a_2a_3a_4\ldots a_m$, where m is any natural number, for the finite decimal that agrees with x out through the m^{th} decimal place and then has all zeros. Also, define $y_m = x_m + 10^{-m}$, so that (by construction of x), for all m, x_m is not an upper bound of S but y_m is an upper bound. Since z∈ S then $z \le y_m$ for all m. Thus we have $x_m \le x < z \le y_m$, for all m. Since, however, $y_m - x_m = 10^{-m}$, which approaches 0 when m approaches infinity. This contradicts the **Squeeze Theorem** we proved before, which says that then x = z. So x is an upper bound for S, as desired.

Now for (2), our construction may yield for x a decimal that ends in all 9's, (apply it, for instance, to the set S={0.9, 0.99, 0.999, ...}), but even if it does, any number smaller than x will be given by a decimal which equals x up to some point and then has a digit which is smaller than the corresponding digit of x. So if y is a smaller number than x, then look at the first digit y has which is smaller than the corresponding digit of x. By construction of x, the decimal which agrees with x out to, and including, this digit is not an upper bound for S, and it is at least as large as y, so that y is not an upper bound, either, as desired. Done with this proof.

Incidentally, with this result, the unique **greatest lower bound** (**g.l.b.**) may also be constructed.

Corollary: Every real number is a l.u.b. or a g.l.b. of a set of rational numbers.

Earlier, to construct the corresponding point on the real line to a given infinite decimal, we assumed the **Cauchy completeness property of the real numbers**:

If [A1, B1], [A2, B2], [A3, B3], ... is any infinite sequence of closed, bounded, nested intervals, whose endpoints are finite decimals, and such that the sequence of their lengths, (Bn-An) with n=1, 2, 3, … , approaches zero, then there is *exactly one* real number which lies in all the intervals.

Now we can prove it as follows.

Proof.

The **Corollary of Squeeze Theorem** shows there is *at most* one such point so we have to prove there is *at least* one. Let S be the set of all left end points of the intervals, i.e. S={An}. Then B1 is an upper bound for S and A1 is in S, so S is not empty and is bounded above, and thus, by **(LUB)**, has a least upper bound, which we call x. Then we claim that x lies in all the intervals. Since x is an upper bound of S, we have $A_n \leq x$, for all n. Moreover, since every Bn is an upper bound for the set S, and x is the least upper bound, no Bn can be less than x, i.e., $x \leq B_n$ for all n. Thus for all n, $A_n \leq x \leq B_n$, i.e., x is in all the intervals. Done.

Importantly, the rational numbers are not Cauchy-complete. For example, the infinite sequence 1, 1.4, 1.41, 1.414, 1.4142, 1.41421, 1.414213, 1.4142135, ... is convergent/Cauchy but it does not converge to a rational number, whereas, in the real numbers, it converges to the positive number whose square equals 2. Nevertheless, to be prudent we should prove that there is a real number whose square is equal to 2. To do this, briefly, we produce an infinite sequence of rational numbers, none of whose squares is equal to 2, but whose squares get closer and closer to 2. We say the real positive square

root of 2 is the real number approximated by our sequence of rational numbers. (Note that we have talked about how to square a real number which is determined by the multiplication operated with the associated sequences of finite decimals.) We will see the formal proof in the next section.

5.2.2 The axiomatic approach

Now, it is time to address the axiomatic approach for the real numbers. If there is a world in which all axioms hold, then those axioms are consistent (i.e., cannot lead to a contradiction) and all theorems deducible from the axioms are true in the world. We may call this world an **axiomatic system**. Therefore, working in an axiomatic system we can take any property as an axiom if we accept it or believe it is true without proof, perhaps for energy-saving, convenience, or simplicity. For example, we can take as axioms the Archimedean properties, LUB, Cauchy Completeness Property of real numbers that we presented before. Moreover, according to Hartshorne (2000, p. 67), if we have set up an axiomatic system, a **model** of that axiomatic system is a realization of the undefined terms in some particular context, such that the axioms are satisfied. For instance, the set of real numbers R is such one model. Hence if we take a model as an example of some world, then we should make sure that the axioms (at least some of the axioms) of the world hold for the model we choose (i.e., models are needed to prove that the axioms we choose are not self-contradictory).

Now let me introduce some pertinent definitions and axioms as follows.

An **ordered field** is a field $(F,+,\cdot)$, together with a subset P whose elements are called “positive”, satisfying

(i) for any $a \in F$, one and only one of the following holds: $a \in P$ or $a = 0$ or $-a \in P$ (the **Trichotomy principle**).

(ii) if $a, b \in P$, then $a + b \in P$ and $a \cdot b \in P$ ("positive" closure);

It follows that $a \neq 0 \in F \Rightarrow a^2 \in P$. This consequence essentially tells us that the product of two negative numbers is positive. We will see the proof in Chapter 6.4.

For example, the rational numbers Q form an ordered field where we take P the positive rational numbers, in the usual sense. Similarly, the real numbers R is also an ordered field with the usual notion of positive numbers. But the field of complex numbers C is *not* an ordered field, since $i^2 = -1$, while $-1 < 0 \notin P$. Moreover, as we talked before Z is only a ring not a field, but with the positive subset N, Z is an **ordered ring**.

Let F be an ordered field. If a set $S \subset F$ and there is an element $x \in F$ such that $x \geq s$ for every $s \in S$, then we say S is **bounded above** and x is called **an upper bound** of S.

In particular, when F be a set of real numbers, and call x an upper bound for F if x is a real number and if x is at least as big as every number belonging to S. Note that if F is the empty set then every number is an upper bound for F.

Let $S \subset F$ be bounded above. We say $x_0 \in F$ is the **least upper bound** (**l.u.b.**) of S if

(i) x is an upper bound of S;

(ii) for all upper bounds x of S, $x_0 \leq x$.

We say F has the **least upper bound property (LUB)** if every nonempty subset $S \subset F$ that is bounded above has a least upper bound.

In particular, when F is a set of real numbers, a least upper bound for F is a number x_0 such that x_0 is an upper bound for F, and no number smaller than x_0 is an upper bound. Note that it is not required that the least upper bound should itself belong to the set F, but only that it shall be a real number. Teachers may want students to reason that the empty set has no least upper bound, and that any set has at most one least upper bound.

An **ordered field** F is called **complete** if and only if F has the **least upper bound property (LUB)**.

The **line-separation axiom** (a corollary of plane-separation axiom):
Any point P on a line L separates the line into two disjoint nonempty convex sides, and given two points A and B different from P on the line L, then A and B are on opposite sides of P $\Leftrightarrow$ P is between A and B $\Leftrightarrow$ P is on the segment AB $\Leftrightarrow$ P is in the interval (A, B) on L if which is the real line. (c.f., Hartshorne, 2000, pp. 74-77)

Dedekind's axiom (D) (the converse of the line-separation axiom):
"If all points of a straight line fall into two classes, such that every point of the first class lies to the left of any point of the second class, then there exists one and only one point which produces this division of all points into two classes, this severing of the straight line into two portions" (Dedekind, 1963, p. 11, as cited in Tiles, 1989, p. 85) (c.f., Hartshorne, 2000, p.115 & p. 139). It says that for every decomposition of the straight line into two nonempty disjoint intervals, exactly one of the intervals has an endpoint. This resolves the contradiction between the continuous nature of the number line and the discrete nature of the numbers themselves. In other words, there is a real number r such

that either one interval equals {x<r or x= r} and the other interval equals {x>r}, or else one interval equals {x<r} and the other equals {x>r or x=r}, with r being the endpoint in both cases. In short, a Dedekind cut yields a real number, which is the reverse of assigning a real number to a given point on a line.

Now I claim the **least upper bound property of real numbers (LUB)** can be considered equivalent to **Dedekind's axiom (D)**.

Proof. (I created this proof with Dr. Roy Smith's direction on March 26, 2010.)

$\Rightarrow$

Assume (LUB), and divide R into two nonempty disjoint intervals J and K. Thus, for R is an ordered field, j<k for every element j of J and every element k of K. And so J is bounded above and every element of K is an upper bound of J (but K may not contain all upper bound of J). By (LUB) we let r be the l.u.b. of J. If J is closed then r is the endpoint of J and r does not belong to K. If J is not closed then r does not belong to J, and so r belongs to K and is the endpoint of K.

$\Leftarrow$

Assume (D), and let R have a nonempty subset S which is bounded above say by b, so every element of S is smaller than b. Then divide R into two disjoint subsets J and K where K consists of all upper bounds for S (thus b is in K), and J consists of all elements which are not upper bounds of S (thus every element of J is smaller than b). Hence, for R is an ordered field, J and K are intervals, since if neither of s, t is an upper bound, so also no points between them are upper bounds and if s, t are both upper bounds then so also are all points between them upper bounds. By (D), we know either one of them has an endpoint r. If r is in J, then K has no left endpoint, and r is not an upper bound for S. Thus

some element x of S is greater than r, i.e., r<x. But r is the greatest element (the right endpoint) of J, so it means x is an upper bound of S (thus x is in K). But by hypothesis x<b. So x is the least upper bound of S, so x is the left endpoint of K. This contradicts to the hypothesis itself. So r is in K not in J, and then it is the least upper bound for S, as desired.

Now let us look at the following three popular axioms for the real numbers:

(1) **the structure axiom**: R is a field;

(2) **the ordering axiom**: R is an ordered set;

(3) **the completeness axiom**: R has a least upper bound property (LUB), i.e., R is Dedekind-complete (D), as we just proved above.

We say R satisfies all the three axioms above, so again as we defined before we call R is a Dedekind-complete ordered field. And so the mapping, f: points on a straight line → real numbers, is a bijection.

Now we can prove the **Archimedean Property (A)**: (algebraic statement) "Given any two positive real numbers x, y, there is an integer *n* such that $ny > x$." as a corollary of **(LUB)** or **(D).** I show it in the following two ways.

Proofs.

(1) Let $S = \{n \in N : ny \le x\}$. If S is empty, i.e., $y > x$, then it is easy to take $n = 1$. Otherwise, when S is not empty, we know that S is bounded above by, for example, $x/y \in R$. Thus by (LUB) or (D), we know there exists a l.u.b. n_0 of S, where $n_0 \in N$ actually is the integer part of x/y plus 1 (if x/y is not an integer). So we can just take n to be $1 + x/y$ such that $ny > x$. Done.

(2) Suppose $n_0 \in N$ is the l.u.b. of N, then $n_0 - 1 < n'$ for some element $n' \in N$, and so $n_0 < (n' + 1 \in N)$, which says n_0 is even not an upper bound of N. So we reach a contradiction, and hence natural numbers are not bounded above, which is what (A) says. (c.f., Birkhoff & MacLane, 1941, p. 69, Theorem 4)

Note that the real numbers R and the rational numbers Q both satisfy the first two axioms (i.e., they both are ordered fields), whereas it is the third axiom that differentiates R from Q. In the aforementioned example (in Chapter 5.2.1) of the approximation rational sequence of the positive square root of 2, all the rational numbers with square less than 2 have a rational upper bound but no rational least upper bound, because the positive square root of 2 is not rational. On the other hand, (LUB) tells us that a real number can be defined by infinite decimals to which the associated Cauchy sequence converges. So we may say Cauchy completeness is a special case of Dedekind completeness. Moreover, certainly (A) does not imply (LUB) or (D) because, for example, the rational numbers satisfy (A), but not (LUB) or (D).

To be circumspect, we now prove the existence of $\sqrt{2}$ as a real number.

Proof.

We want to prove that $\sqrt{2}$ exists as the l.u.b. of the set $S = \{s \in Q : s^2 < 2\}$. It is obvious that S is bounded above (say, by 2), so by (LUB) or (D) it has least upper bound denoted by l. Hence we want to prove $l = \sqrt{2}$, so to prove $l^2 = 2$.

By the **Trichotomy principle** we now prove that $l^2 < 2$ and $l^2 > 2$ both lead to contradictions and so we must have $l^2 = 2$.

Thus we suppose $l^2 < 2$. We now consider $\left(l + \frac{1}{n}\right)^2 < 2$ with $n \in N$, which is equivalent to $l^2 + \frac{2l}{n} + \frac{1}{n^2} < 2$. Then we have $l^2 + \frac{2l}{n} < 2$, and so $n > \frac{2l}{2 - l^2}$. Then by (A) we can find such an n so that $\left(l + \frac{1}{n}\right)^2 < 2$. But $\left(l + \frac{1}{n}\right) > l$, so l would not be an upper bound. This contradicts the hypothesis l is l.u.b. So $l^2 \not< 2$, as desired.

Similarly, we suppose $l^2 > 2$. We now consider $\left(l - \frac{1}{n}\right)^2 > 2$ with $n \in N$, which is equivalent to $l^2 - \frac{2l}{n} + \frac{1}{n^2} > 2$. Then we have $l^2 - \frac{2l}{n} > 2$, and so $n > \frac{2l}{l^2 - 2}$. Then by (A) we can find such an n so that $\left(l - \frac{1}{n}\right)^2 > 2$. Since $\left(l - \frac{1}{n}\right) < l$, we found another upper bound which is smaller than l. But this contradicts the hypothesis l is l.u.b. So $l^2 \not> 2$, as desired.

Done with the proof.

In short, Dedekind's axiom holds in an ordered field if and only if it is a complete ordered field. Any two complete ordered fields are isomorphic; there is one and (except for isomorphic fields) only one complete ordered field, which actually is R (Birkhoff & MacLane, 1941, pp. 71-73). These are essentially the same as Harthorne (2000, p. 70) talked which we will briefly mention later in Chapter 5.4. The proofs for these conclusions are long and not easy, and would lead us far afield, so I just mention them here for people who have further interest.

5.3 Uncountability/countability and polynomials

We have known that natural numbers and any subset of it are countable. What about real numbers? How many points are there in a line? We claim that there really are more real numbers than there are natural numbers, namely, positive integers, in a rather precise sense.

Theorem: (Georg Cantor) The set of real numbers is not countable.

Proof (adapted from Dr. Smith's Paideia notes plus my own interpretation).

We will just show that no list of real numbers can contain all of them. Indeed we will restrict ourselves just to those real numbers lying between 0 and 1 and whose decimal expansion contains only 0's and 1's. What we will do is assume that we are given a list of such real numbers and then give a way of cooking up another such real that cannot possibly be in the list. That will show that the list cannot be complete.

So let us take an example of one possible list:

1) 0.0000000000...

2) 0.1000000000...

3) 0.1100000000...

4) 0.1110000000...

5) 0.1111000000...

……

Now of course it is easy to see that 0.0101010101010101... , for example, is nowhere in this list, but let us try to construct another number, not in the list, in a more systematic way, a way that has some hope of working on all other lists as well.

Notice first of all that in order for two decimals of this type to be different they only need to be different in one digit. So, for instance, to construct a decimal that is different from the first one in this list, we only need to let its decimal part begin with a 1, instead of a 0. So we let our number start out as 0.1, and then we have already made it different from the first number in our list, and we are nowhere near finished constructing it. How should we continue it so that it will be different also from the second number in the list? Since we are ready to choose the second digit of the decimal part, and since the second number in the list has 0 as the second digit of the decimal part, we only need to choose the second digit of the decimal part of the third number as 1. So our decimal now looks like 0.11, so far. Now look at the third digit of the decimal part of the third number in the list, which is again a 0, and conclude therefore that we should choose a 1, and we have 0.111, for our number so far. Get the idea? We are led to the number 0.111111... , which is different from every number in the list.

Now let us try it in general.

Suppose we are given some infinite list of infinite decimals like those above:

1) $0.a_1a_2a_3a_4a_5...$

2) $0.b_1b_2b_3b_4b_5...$

3) $0.c_1c_2c_3c_4c_5...$

……,

where each of these digits is still each 0 or 1.

Now to construct an infinite decimal which is not in the list, we do as we just did. To figure out a way to write it, let us make a rule. We put a "¬" symbol over a digit to

change that digit into the opposite choice of digit, 0 or 1. Thus if a_1 is 0 then $\bar{a}_1$ is 1, and vice versa. Then our number, which we claim is nowhere in the list, can be written as $0.\bar{a}_1\bar{b}_2\bar{c}_3\cdots$. That is, our number has a different first decimal digit from the first number in the list, so it does not equal the first number. It also has a different decimal second digit from the second number in the list, so it does not equal that second number either. Indeed our number cannot equal any number in the list since it differs from each number in at least one digit. Thus we have constructed a real number whose decimal expansion has only 0's and 1's and which is not in the given list. Thus no one list can contain all real numbers of that type. That is, the real numbers of that type are "**uncountable**". This implies also that the "larger" set of all real numbers is uncountable too (since the real numbers in [0, 1] have a one-to-one correspondence with the numbers in any other interval of unit length on the real line). Done.

We have kept using the concept "polynomials" above. I think it is the time to define it formally as follows, and then I give my pedagogical concern which is related to this concept.

Definition (Birkhoff & MacLane, 1941, p. 77): Let D be any **integral domain** [why not any ring? – we will see in a little while one possible reason associated with the degrees of polynomials], 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 (technically, the postulates for an integral domain). This procedure is, as least in special cases, familiar from high school algebra, and leads to the construction of

polynomials in x. Since nothing is assumed known about x (not even that it is an unknown element of D), x is usually called an **indeterminate**.

Generally, the starting exploration of the algebraic structure for polynomials can be put parallel to that for natural numbers. The analogy of Division Algorithm for both natural numbers and polynomials is an example; the analogue of prime numbers is irreducible polynomials, and so is the analogy of the unique factorization of a natural number and that of a polynomial in a given field. It is worth noting at this point that, however, this definition treats polynomials for their own sake, different from what most high school students' understanding that a polynomial is an expression with values or a function of a variable on a set of numbers, i.e., the indeterminate x in a polynomial needs not be a variable, and so computation with polynomials is not the computation of the evaluations of the polynomials "at" certain numbers, hence, so do rational functions (quotients of polynomials) or even rational expressions (quotients of algebraic expressions, including polynomials). Therefore, I do not quite like the routine that the teacher suggests students checking the validity of an algebraic transformation by substituting numbers, because then the students would be unfortunately unconsciously reduced to only working in the field of numbers or even worse only exercising number manipulation instead of thinking on the standpoint of the field of rational expressions; it is harmful for students to develop a conscious interpretation to algebraic expressions as mathematical objects in their own right. Thus, for example, when students get the algebraic identity $\frac{x^2-25}{5-x}=-(x+5)$, the teacher should be sensitive to the context the students are asked to work in and careful about saying "as long as x does not equal 5" or "except when x equals 5. The same reason works for the equation

$1+x+x^2+...+x^{p\text{-}1}=\frac{x^p\text{-}1}{x\text{-}1}$ we saw in Chapter 3. Given this, it then would not be very difficult for students to develop the crucial notion of substituting an algebraic expression for a variable in a function; one of the typical questions is that if $f(x)=x^2-2x+5$, then what is $f(x+\sqrt{2})$?

If we write a **polynomial** with coefficients in a commutative ring r in the form of $f(x)=a_nx^n+a_{n-1}x^{n-1}+...+a_1x^1+a_0$ where $a_i\in r\,|_{i=1,2,...,n\in N}$ and $a_n\neq 0$, i.e., $f(x)\in r[x]$, then an individual term of the form a_ix^i is called a **monomial**. The exponent of the monomial with the highest exponent is the **degree** of the polynomial, denoted by $\deg(f(x))$, and then the coefficient of that monomial is called the **leading coefficient** of the polynomial. Teachers may let students check why $r[x]$ is a commutative ring when r is a commutative ring (for the proof, see Shifrin, 1996, p. 83, Proposition 1.1). We call $s\in r$ is a **root** of $f(x)\in r[x]$ if $f(s)=0$.

Incidentally, we can have polynomials also with coefficients in a non-commutative ring, but we have to be more careful about multiplying them. And it is not as easy to substitute values into them; we have to substitute either from the right or from the left (R. Smith, personal communication, December 30, 2009).

We, however, claim that we can find a subset of real numbers which is a countable set, i.e., all the real numbers contained in the set of all **algebraic numbers**, defined as the (complex) numbers which are solutions of some non-zero polynomial equation $a_nx^n+a_{n-1}x^{n-1}+...+a_1x^1+a_0=0$ with integer (or, equivalently rational) coefficients and positive integer exponents.

Proof.

Let us say that, for a given non-zero polynomial equation with integer coefficients, we will call the "size" M (a natural number) of the equation the integer we get by adding together the degree of the equation and the absolute values of all the coefficients. For example, the size of $2x^3 + 17x - 5 = 0$, is N=3+(2+17+|-5|)=27. Then let the set S_1 be the set of all equations of size 2 (why not starting with size 0 or 1?), and S_2 be the set of all equations of size 3, and so on. Then I **claim** that:

(1) there are only a finite number of equations in each set;

(2) any non-zero polynomial equation has only a finite number of solutions.

Now we can make a list of all algebraic numbers by taking all the solutions of equations in the set S_1 and putting them in the list first, one after another until we have all of them. Then after those, we can put down all the solutions to the equations in the set S_2, and so on. In this way we will eventually list all solutions of every equation, i.e., we will be able to list every algebraic number, so that we have proven that the algebraic numbers are countable.

Now we get an interesting bonus! Since the algebraic numbers are merely countably infinite, whereas there is a much larger, uncountable infinity of all real numbers, it follows that some real numbers (most of them) are not algebraic!

But I have leave out the proofs of claims (1) and (2) in the above proof. Now I will prove them as follows.

Proof for claim (1) by induction:

Let us first check the first two cases.

We do not consider the case M=0 and M=1, since they would turn out to be zero polynomials, which look like $0x^0 = 0$ and $0x^1 = 0$, which have infinite solutions, and the degree of which actually is undefined.

When M=2, $S_1 = \{(\pm 1)x^1 = 0\}$, so there are three elements in this set, which is a finite one.

When M=3, $S_2 = \{(\pm 1)x^2 = 0, (\pm 2)x^1 = 0, (\pm 1)x^1 \pm 1 = 0\}$, which is a finite set with 9 elements.

Now suppose S_N is a finite set when $2 \le M \le k$, where $k \in N$.

We want to prove S_{k+1} is also a finite set. Let us consider all the possible polynomials of S_{k+1}. $1x^k = 0$ must be one of them. It is also the only possibility for the highest degree k. So the other elements must be of degree less than k. But we supposed that S_k, S_{k-1}, … , S_1 are all finite sets, so S_{k+1} is also finite, as desired. Done.

Let us restate **claim (2)** more strictly as an algebraic **proposition (*)**:

A polynomial of degree n with coefficients in a field F has *at most* n different roots in F.

Thus, we have interest in the ring $F[x]$, namely, the ring of all the polynomials with coefficients in the field F, and in this case the following **Division Algorithm of polynomials** (like the Division Algorithm we introduced before for the ring of natural numbers) holds.

Let $f(x), g(x) \in F[x]$ be non-zero polynomials. Then there are unique polynomials $q(x), r(x) \in F[x]$ such that $f(x) = q(x)g(x) + r(x)$, with $\deg(r(x)) < \deg(g(x)) \leq \deg(f(x))$ or $r(x) = 0$.

Note that this is the precise description of "long division" of polynomials usually taught in high school algebra.

Lemma: Let r be an integral domain. If $f(x), g(x) \in r[x]$ are non-zero polynomials, then $\deg(f(x) \cdot g(x)) = \deg(f(x)) + \deg(g(x))$. Moreover, $r[x]$ is an integral domain. By the way, this property of degree may explain why Birkhoff and MacLane (1941) defined polynomials with coefficients in an integral domain instead of any non-commutitative ring, i.e., he might want to have some better properties, more like integers, for his polynomials.

Proof. Let $\deg(f(x)) = n$ and $\deg(g(x)) = m$. Since r is an integral domain, the leading coefficients a_n and b_m are not zero-divisors, and so $a_n b_m \neq 0$. Thus, $\deg(f(x) \cdot g(x)) = n + m$. Hence, $\deg(f(x) \cdot g(x)) = \deg(f(x)) + \deg(g(x))$. In particular, if neither $f(x)$ nor $g(x)$ is the zero polynomial, the product $f(x) \cdot g(x)$ cannot be the zero polynomial, so $r[x]$ is an integral domain. Done.

Of course, this conclusion holds for r is a field, i.e., $F[x]$ is an integral domain when F is a field.

Remainder Theorem (corollary of the Division Algorithm):

Let $s \in F$ and $f(x) \in F[x]$. When we divide $f(x)$ by the monomial $x - s$ the remainder is a constant $f(s) = c$.

Proof.

By the **Division Algorithm**, we have $f(x)=(x-s)q(x)+r(x)$ (*), where $\deg(r(x))<\deg((x-s))$ or $r(x)=0$. But $\deg((x-s))=1$, thus $r(x)$ is a constant c no matter what value x takes, which could be 0. Take $x=s$ in (*), then we get $f(s)=r(s)=c$. Done.

Referring to the definition of **a root of a polynomial** (defined above), when $r(x)=c=0$ we get $f(x)=(x-s)q(x)$, and so $f(s)=0$, thus, $s\in F$ is a root of $f(x)\in F[x]$. Hence, we get an important consequence, namely the **Root-Factor Theorem**:

Let $s\in F$ and $f(x)\in F[x]$. Then $x-s$ is a factor of $f(x)$ if and only if s is a root of $f(x)$, in symbols, $(x-s)\mid f(x)\Leftrightarrow f(s)=0$.

Notably, as mentioned in Chapter 3, this theorem as our **second irreducibility criterion** explains why some polynomials $f(x)\in Z[x]$ such as x^2+2 and x^3-32 are irreducible in $Q[x]$ while they do not satisfy Eisenstein's Criterion.

Now we prove the above **proposition (*)** as follows.

Proof.

Let $\deg(f(x))=n\in N$. By the **Root-Factor Theorem**, if $a_1\in F$ is a factor of $f(x)\in F[x]$, then $f(x)=(x-a_1)f_1(x)$, where $\deg(f_1(x))=n-1\in N$. Then if $a_2\in F$, but different from a_1, is a factor of $f(x)\in F[x]$, then $f(a_2)=(a_2-a_1)f_1(a_2)=0$, where $a_2-a_1\neq 0$, so $f_1(a_2)=0$. And so $f_1(x)=(x-a_2)f_2(x)$, where $\deg(f_2(x))=n-2\in N$.

Thus, $f(x)=(x-a_1)(x-a_2)f_2(x)$. By **induction**, suppose $a_1,a_2,...,a_m \in F$ are all distinct roots of $f(x)$, then $f(x)=(x-a_1)(x-a_2)...(x-a_m)$, in other words, each root gives rise to a linear factor. By the **Lemma** proved above,

$$\deg\left((x-a_1)(x-a_2)...(x-a_m)\right)=\deg\left((x-a_1)\right)+\deg\left((x-a_2)\right)+...+\deg\left((x-a_m)\right)=m.$$

But $\deg\left(f(x)\right)=\deg\left((x-a_1)(x-a_2)...(x-a_m)\right)=n$, so $m\le n$, in other words, $f(x)$ cannot have more than n linear factor. Therefore, a polynomial of degree n with coefficients in a field F has *at most* n different roots in F. Done.

Here is another perspective to state this proposition, i.e., the number of different roots of a polynomial is not greater that the degree of the polynomial. Therefore, we can also start from assuming that if there are at least n different roots of the polynomial $f(x)$. Thus, still by the **Root-Factor Theorem** and **induction** we may get

$f(x)=\prod_{i=1}^{n}(x-a_i)g(x)$, where $\deg\left(g(x)\right)\ge 0$. Therefore, also by the **Lemma** $\deg\left(f(x)\right)\ge\left(n+\deg\left(g(x)\right)\right)\ge n$, that is, the number of different roots of $f(x)$ is equal to or smaller than the degree of $f(x)$, as desired.

Till now have we finished the proof that algebraic numbers are countable by going through above basic concepts and theorems about polynomials, which I think high school algebra should give more weight to.

It looks like square root of a rational number is not as usual as rational numbers, but we still can get a rational number back when we square it. So again there is no doubt that square root of a rational number is an algebraic number. But how do we know the square root of a positive rational number is not rational? (There should be no doubt that the

square root of a negative rational number or an irrational number is not rational.) I will prove $\sqrt{2}$ is an irrational number as a classic example in two different ways as follows.

Proof by contradiction.

Suppose $\sqrt{2}$ is a rational number, then let $\sqrt{2} = \frac{n}{m}$, where *n* and *m* are the "positive" elements of the **ordered ring** Z. So we have $2 = \frac{n^2}{m^2}$, and so $n^2 = 2m^2$. But in the prime factorization of n^2 there are twice as many 2's as in the factorization of *n*, hence an even number, and the same holds for m^2. But that means in the prime factorization of n^2 there occur an even number of factors of 2, while in the factorization of $2m^2$ there occur an odd number (the even number of factors of 2 in m^2, plus the extra "2" in front of $2m^2$). Since an integer has one **unique prime factorization** (c.f., Chapter 5.1) (or in terms of abstract algebra, the integers is a unique factorization domain (**UFD**)), it cannot be true that $n^2 = 2m^2$. Therefore, $\sqrt{2}$ is not a rational number. Done.

By the way, teachers may want students to think whether the proof for $\sqrt{3}$ would be different. Note that 3 is not even, the popular high school proof (different from the one stated above), which assumes the "lowest term" and then using $n^2 = 2m^2$ reaches a contradiction that the numerator and the denominator of the lowest term are both even numbers, might need some changes. What about the case for square root of a non-prime number? This way of proving might have its deficiency.

There is another proof which involves polynomials not just manipulation of numbers, by using the **Rational Root Theorem**:

Given a primitive polynomial $f(x) = a_n x^n + a_{n-1} x^{n-1} + ... + a_1 x^1 + a_0 \in Z[x]$ (i.e., all the coefficients of a polynomial with integer coefficients have no common integer divisor except for ± 1), if $\frac{s}{t}$ is a rational root of $f(x)$, where $s, t \in Z$ and $\gcd(s,t) = 1$ (i.e., $\frac{s}{t}$ is expressed in lowest terms), then $s \mid a_0$ and $t \mid a_n$.

The proof is not complex. Teachers may want students to try it out. By the way, this theorem is very useful in calculus in maximizing polynomials by finding zeroes of their derivatives.

The usual practice or application of this theorem in high school stops at those questions like "List all possible rational zeros of the each function." or "Find the possible rational solutions of the following polynomial equation.". Actually it directly implies the fact that square root of 2 is an irrational number, instead of scaring students by not telling them the immediate corollary. And applying this theorem we will know whether a primitive polynomial with integer coefficients is irreducible over the rational numbers, so this is our **third irreducibility criterion** different from Eisenstein's Criterion which we introduced in Chapter 3.

What is a square root of a (real) number? I got, from some mathematicians and graduates of mathematics and mathematics education at UGA, the most frequent explanation: the square root of a number x is a number y so that y times y equals x; it may or may not exist in the number system we were working in. This is consistent with the statement that the square root of a rational number is an algebraic number. So we say $\sqrt{2}$ is a positive root of $x^2 - 2 = 0$. Now we want to check whether it is a rational root.

Suppose $\frac{s}{t}$ in lowest terms is a rational root of $x^2 - 2 = 0$. Thus, by the **Rational Root Theorem**, we have $s \mid (-2)$ and $t \mid 1$, and so all the possible rational roots are ± 1, ± 2. Remember the theorem itself does not guarantee there is a rational root, so when we plug them back in $x^2 - 2 = 0$ we find they are even not its roots at all. Hence we conclude that $\sqrt{2}$ as a root of the polynomial just will not be rational. Note that if we draw the graph of $y = x^2 - 2$ on the Cartesian plane over the rational numbers, out of most high school students' expectation, we will not see the two intersections of the graph and the *x*-axis, both of which are consist of only rational numbers. Here is an alternative approach. Supposing ± 1, ± 2 are all rational roots of $x^2 - 2 = 0$, we can check $\sqrt{2}$ is not equal to none of them. Note that some students probably will say it is obvious since, for example, we get $\sqrt{2}$ by taking positive square root of 2 while 2 is itself without being operated. So we thought "How could they be the same?". I once made this conceptual mistake too. But I was questioned about how I know $\sqrt{2} \neq 2$ just because they look different (R. Smith, personal communication, December 16, 2009). Therefore, the thing turns out to be finding out when $\sqrt{\alpha} = \alpha$ with $\alpha \in R$ for example. In other words, when $\alpha = \alpha^2$? If we draw the graphs of the functions $y = x$ and $y = x^2$ on the Cartesian plane over R, as shown in the figure below, we will see that the intersections of the two graphs are (0,0) and (1,1), i.e., only when $\alpha = 0$ or 1 we have $\sqrt{\alpha} = \alpha$. So we conclude $\sqrt{2} \neq 2$.

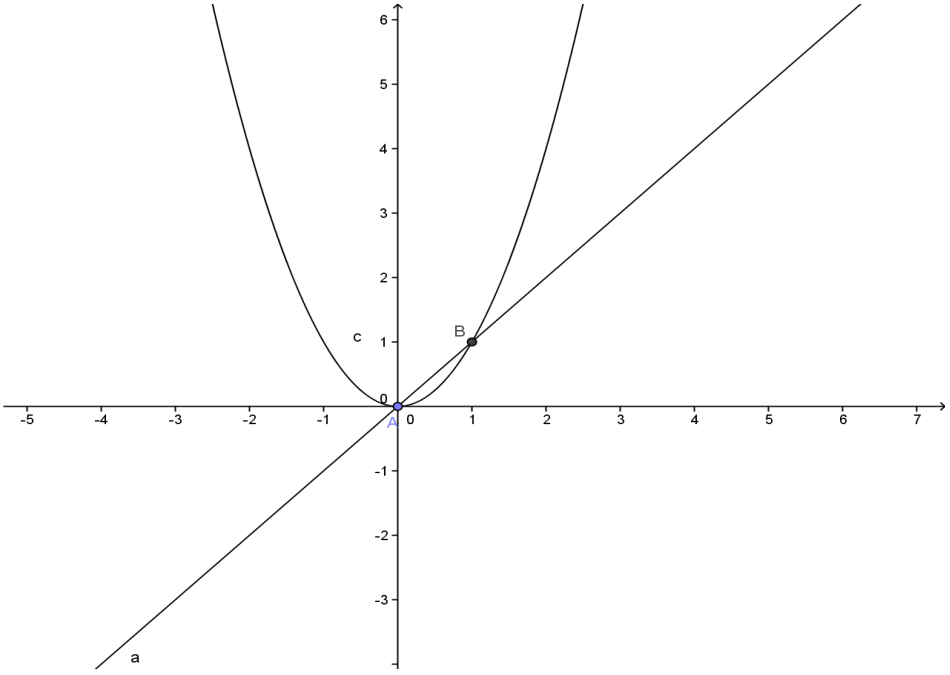


Figure 5.3.1 Intersections of the graphs of functions $y = x$ and $y = x^2$ on R^2

5.4 Algebra and analytic geometry

I believe real numbers virtually grow out of geometry. We could see this when we were trying to assign a real number to a point on a line. Another origin might be related to Pythagorean Theorem, which tells people that the diagonal of a unit square (a geometric object) is a number whose square (an operation) is double unit, say 2, if we take the unit to be 1. People realized that they could not find a natural number or even a rational number whose square is 2. If the ratios of two pairs of segments are equal in our usual sense about real numbers, we may want to see what will happen if the two pairs of segments are two pairs of sides of two triangles. Hartshorne (2000) commented that while Euclid was able to develop the material of book I~IV of the *Elements* without any notion of number, it is a different matter when we come to the concept of similar triangles as taught in high school. These are triangles whose sides are not equal, but all in some

common “ratio” to each other. If that ratio is an integer, then we see the length of the longer side is multiple of the shorter one; actually, the shorter one can be consider as a unit copy of the longer one. If that ratio is not an integer but a rational number, i.e., the sides of the similar triangles are integer or even rational multiples of each other. In other words, if we impose the vertex of the smaller triangle on the vertex of the bigger one, and impose their sides along each other, by rigid motions (assuming its existence), then the lines where their bases lie are parallel, and If the ratio is $\frac{n}{m}$ with $n, m \in Z$ and $\frac{n}{m} < 1$, then we will see that there are exactly n subdivisions of side of the smaller triangle lie above its base and (m-n) the same subdivisions below its base. Euclid loved rational numbers. But he noticed that he could not always find such subdivision for two “similar” triangles, actually even for the division of a pair of segments. But there are always same numbers of subdivisions lie above and below the “base line” of the smaller triangle on the two sides of the bigger triangle. So Euclid decided that he needed irrational numbers to express the ratios (c.f., Euclid’s *Elements*, Definition V.5 (equivalent to our description about how the real numbers behave, say that two real numbers are equal if and only if the two real numbers are both greater than or equal to the same rational numbers), Definition V.6, & Definition VI.1).

I think it is also good for high school students to know the following information. When Euclid constructed a segment he used compass to construct a circle which intersects with a straight line, and he actually was assuming **Archimedean axiom** and the **existence of the circle-circle intersections** (equivalently, the **existence of the circle-line intersections** (c.f., Hartshorne, 2000, p. 144, Proposition 16.2) (both of which indeed can be implied by **Dedekind’s Axiom**). (By the way, Hartshorne (2000, Ch.4) showed that he

could prove Euclid's propositions *without assuming Archimedean axiom* by creating an arithmetic of line segments.) We all know in the real Cartesian plane the equation of a circle is a quadratic one and the equation of a line is a linear one. So the intersections of one circle and one segment are the common solutions of the two equations of them, and so this kind of construction actually only allowed us to get a segment of a special irrational length which is (a multiple of) a square root of a positive integer. Hence, when Euclid used ratio to compare two segments, he also got only a special real number which is (a multiple of) a square root of a positive rational number. Hence, modern mathematicians realized that Euclid's real numbers, which are actually what high school geometry are talking about, are not all the real numbers on the real line. An **ordered field** F is defined to be **Euclidean** if and only if $x \in P \Rightarrow (\exists y \in F)(y^2 = x)$. So all ratios of pairs of segments in Euclidean geometry are the positive elements of a Euclidean field. We define the Cartesian plane over a field F as $F^2 = F \times F = \{\text{all ordered pairs } (x, y) \text{ with } x, y \in F\}$, and a line as a subset of F^2 is defined as the solution set of a linear equation: $\{(x, y) \in F^2 : ax + by = c \text{ with } a, b, c \in F \text{ and } a, b \text{ not both zero}\}$. Consequently, like real numbers R, if F is any Euclidean field, the Cartesian plane over F turns out to be a Euclidean plan in which all Euclid's propositions hold. Algebraically, the elements of our Euclidean field, which essentially defines high school geometry, are some elements of R which are obtained by a finite number of operations of $\{+, -, \times, \div, \sqrt{\ }\}$ on rational numbers Q. So the Cartesian plane R^2 is a model for Euclidean geometry, but is a bit "overloaded". Note that **Dedekind's axiom (D)** holds if and only if Euclidean numbers

are real numbers. In other words, a **Euclidean plane** (a set of "points" with subsets called "lines")) is defined to be a **Hilbert plane** (assuming the entire list of **Hilbert's axioms** (Incidence; Betweenness; Congruence) satisfying the **circle-line or circle-circle intersection axiom**, and **Playfair's axiom**, also called the **parallel axiom**, which says for a line L and a point P not on L, there is at most one line containing P which is parallel to L. Note that a Euclidean plane does not have to satisfy **Archimedean axiom**, but I prefer to assume it also, which makes our axiom system more natural. So we say a **Euclidean plane** with **(D)** is isomorphic to the **Cartesian plane over the field of the real numbers**. In other words, putting Hilbert's axioms, Playfair's axiom, and Dedekind's axiom, together, the axiomatic system will be categorical, i.e., the **unique** model (up to isomorphism) for the system will be the real Cartesian plane (Hartshorne, 2000, p. 70). Last, please allow me to be a little bit picky. Similar to the side-effect of calculators and graphing software I mentioned before, GSP is good for students to explore some geometry figures, but people may not notice that it does not really a model of Euclidean geometry, since all the numbers used in it have to be rounded to rational numbers.

Theorem (Birkhoff & MacLane, 1941, p. 67): If $p(x)$ is a polynomial with real coefficients, if $a < b$, and if $p(a) < p(b)$, then for every constant *c* satisfying $p(a) < c < p(b)$, the equation $p(x) = c$ has a root between a and b.

Birkhoff and MacLane (1941) commented that "[g]eometrically, the hypothesis means that the graph of $y = p(x)$ meets the horizontal line $y = p(a)$ at $x = a$ and the line $y = p(b)$ at $x = b$; the conclusion asserts that the graph must also meet each intermediate horizontal line $y = c$ at some point with an x-coordinate between a and b" (p. 67), and

"there is a general theorem of analysis which asserts this conclusion, not only for polynomial functions $p(x)$, but for any continuous function" (p. 67 in the footnote). Birkhoff and MacLane also proposed that this theorem "does not show how to construct numerical solution to numerical equations; it merely proves their existence" (p. 69), and the study of computing them "is not a part of algebra, but of analysis" (p. 69). So here we may assume that we can add, subtract, multiply, and divide real numbers as assumed in high school algebra (i.e., we assume R is a field).

Note that the theorem (Birkhoff & MacLane, 1941, p. 67) above actually is the **Intermediate Value Theorem** in calculus, which can be considered to be equivalent to **Dedekind's axiom (D)**, claimed by Dr. Roy Smith (personal communication, Fall 2009), while **(D)** reflects the modern development of the real numbers and notions of continuity.

6. Reflections and Suggestions

"All mathematics students should be exposed to the basic ideas of modern algebra, its problem-solving skills and basic proof techniques, and certainly to some of its elegant applications" (Shifrin, 1996, p. vii), but do not forget that students learn what they are taught which can serve to broaden or limit their view or understanding on algebra. The picture I have painted so far is an imperfect, tentative, but hopeful one, with a goal of giving some of my own ideas about helping high school students and teachers develop a more positive and structural view about algebra and mathematical thinking and an appreciation of the beauty of mathematics. Additionally, I believe that the process itself of writing a thesis is a process of learning, and that this thesis will be a very precious document for my own teaching in the future.

6.1 Mathematics itself worth teachers' and students' appreciation

Many have noted the beauty of mathematics. For instance, King (1992) noted that "One of the vastest areas of the world of contemplative beauty is mathematics. This alone is sufficient reason for its study" (pp. 275-276). And Lang (1985b) said "whereas the beauty of poetry pales under translation, the beauty of mathematics is invariant under linguistic translations" (p. 18).

King (1992) said that nobody began with mathematics because of its beauty; all school children began the study of mathematics because they had no choice. Some

children just fell in love with mathematics from the beginning, or at least as soon as they found out they were good at it, while some others just quit as soon as they were done with school. There are also some children who endured, wanting to see the great practical value emphasized by their teachers until they, entirely by chance, encountered pure mathematics and felt like breathing the fresh air on the deck after living all their lives in the hold of some great ship (King, 1992). However, King (1992) said that his early teachers over-chanted the notion of practical value of mathematics by justifying mathematics on the basis of its utility in the conduct of one's daily life. The true value of mathematics lies outside of pedestrian commonplace activity. The intrinsic worth of mathematics itself is a creative and intellectual art, and the value stemming from mathematics is its unreasonable effectiveness in explaining and predicting real-world, physical phenomena; to fully appreciate either of these values, one must seriously study mathematics to some level, a clear cut of which probably does not exist (King, 1992). Similarly, Saul (2008) proposed a broader view of real life as "the life of the mind" (p. 75) and noted that "[m]athematics itself acquires a reality for students, … motivation for learning mathematics can eventually arise from questions within mathematics itself. That is, mathematics is, at a certain level, its own application" (Saul, 2008, p. 75). I also believe that mathematics taught as "low art" (King, 1992, p. 277) is definitely not enough. King (1992) said that students come to college, memorizing the mathematics "told" (p. 276) to them, "with no feel for mathematics that can, in any way, be associated with art. In fact, they seem convinced that either no such sensation is possible, or else that it falls – like ultrasonic sound – outside the sensory range of ordinary mortals" (p. 277). King

(1992) tried to make clear that he did “not believe that a feel for mathematics is innate” (p. 277) and commented that

> The fact that mathematics presently lies outside the artistic range of most people is the fault of neither the audience nor of mathematics. What has gone wrong is the manner of presentation. How else can there exist a person who likes poetry and hates mathematics? Properly presented, they are much the same. (p. 277)

Algebra should be viewed as a body of knowledge and also as a way of thinking; so is mathematics, of course. Students should be encouraged to engage in reflecting on the meanings of algebraic expressions and also constructing meaningful ones while discussing with their teachers and peers, instead of doing an excessive amount of time-wasting homework that just requires practicing rudimentary skills of manipulating. "Not every student needs proficiency in symbol manipulation skills. By choice or circumstance, many students will never reach the levels of mathematics study where they will use these skills. However, every student needs to understand how quantities depend on one another, how a change in one quantity affects the other, and how to make decisions based on these relationships” (Williams & Molina, 1998, p. 41). This reminds me of an old Chinese saying "Facing what you are seeing, you should know how and why it is that."

Even though his students were successful on standardized tests, Chazan (2000, p. xiv-xv) was skeptical that his students had learned much of lasting value about algebra and was uncomfortable that they did not exercise their own judgment in mathematics. And he once was unhappy with his own understanding of algebra as subject matter, which needs to be fundamentally different from that which had created daunting and debilitating experiences for students.

In order to emphasize what I meant to say, I quote King’s (1992) vivid description (it is sad but not at all universal):

> All of us have endured a certain amount of classroom mathematics. We lasted, not because we believed mathematics worthwhile, nor because … we found the environment favorable. We endured because there was no other choice. Long ago someone had decided for us that mathematics was important for us to know and had concluded that, if the choice was ours, we would choose not to learn it. So we were compelled into a secondary school classroom fronted with grey chalkboards and spread with hard seats. A teacher who had himself once been compelled to his same place stood before us and day after day poured over us what he believed to be mathematics as ceaselessly as a sea pours forth foam. The room in which we sat was a dark and oppressive chamber and … only in a fallen world could such a place exist. (pp. 15-16)

King (1992) then identified three groups of people in the classroom where students (a group of "scientists" and a group of "humanists") had precollege mathematics thrust upon them. Unfortunately, the high school mathematics teacher (as the only member of the third group) had three characteristics: "he did not like mathematics, he did not understand mathematics, he did not believe mathematics important" (p. 16). King (1992) explained his opinions as follows. First, the teacher did not like mathematics was clear from the beginning.

> His lack of fancy for the subject he taught came to us … from his transparent absence of passion for mathematics. … Passion, we knew even then, is too easily communicated. But from our teacher we heard nothing of mathematics except basic facts. When he spoke to us of mathematics he spoke with neither ardor nor metaphor. He taught mathematics on weekdays with less enthusiasm than he showed on Saturday when he mowed his lawn. We knew he did not like mathematics. But we did not hold that against him. We did not like it either. (pp. 16-17)

Then later slowly after the students themselves had studied more advanced mathematics they realized that

> [w]e could then –as earlier misunderstood notions became clear– look back on what he had told us about mathematics and pinpoint exactly the shallowness of his understanding. But we saw unmistakable signs of his ignorance even as he taught us. Mostly [he] showed through his fumbling and fearful responses to elementary questions that he could not answer. (p. 17)

King (1992) made this “understandable” by telling the “truth” that at that stage in one’s life neither the students nor the teacher expect anyone to understand mathematics. It is just something they endure for as long as they must. Consequently, it becomes “understandable” also to the students that the person who teaches them mathematics in high school does not think it is valuable.

> Why should he? No one else you know does. Your parents live their lives without mathematics and so do your parents’ friends. … Mathematics is mentioned neither in the newspaper nor on television. At no time has mathematics ever been – within your range of hearing – a subject of conversation.
>
> Naturally, your teacher *tells* you almost every day that mathematics has value. But you know that he does not believe it. And he knows that you know. It is just another shared fiction. … (p. 18)

It is so unpleasant to see “this unhappy band of three” (King, 1992, p. 18) went on.

> … [Students] took a succession of secondary school courses. And, at each stage, the separation between the future scientists and the future humanists became more and more sharply defined. The “scientists” … determined that while the subject might never be understood, it could at least be learned. As the curriculum advance, the “humanists” became more and more ignored by the mathematics teachers and advisors, and were allowed -- even encouraged -- to drop out of the mathematics sequence.

The mathematics teachers, as might be expected, continued. They came before us one by one, uninspired and uninspiring, as identical as dominoes. (King, 1992, pp. 18-19)

6.2 My pedagogical understanding of mathematical thinking

I guess the concepts and facts in high school algebra were generated and extracted from abstract algebra so that high school students can understand these ideas. While there are many more concepts in high school algebra that are connected to abstract algebra, I have been working on exploring a limited number of connections that I have seen as a master’s student majoring in mathematics education. I have been using a less sophisticated way to illuminate those connections that I think important for high school

students to know so as to enrich their perspectives on the algebra they learn in high schools. To be clear, I am *not* saying the way I presented the content is the most efficient way to help teachers and students to see some algebra concepts taught in high school as connected to abstract algebra. In fact, "efficiency is often in the eye of the beholder when it comes to mathematical thinking" (Boaler & Humphreys, 2005, p. 16). The same task could be implemented in a number of different ways and at a number of different levels, which will also affect the cognitive effort students will put into the learning and the mathematical connections they will see.

"To engage students in challenging mathematics, teachers need to use worthwhile mathematical tasks. Those tasks must be rich in terms of content and processes" (Martin, 2007, p. 39). In this thesis I concentrated upon describing the connections I noticed in an explicit way with much less focus on the task processes. In addition, providing a prescriptive list of things to do to implement mathematical tasks designed to help students see connections becomes difficult due to the dynamic aspects of the classroom. But I suggest that teachers put more effort in designing and implementing worthwhile mathematical tasks that would engage students in high-level cognitive and intellectual thinking.

> Tasks that promote communication and connections can help students see and articulate the value and beauty of mathematics. Teachers can enhance the value of existing materials by tailoring them to needs and interests of their students. By doing so, teachers can promote both motivation and equity in their classrooms. (Martin, 2007, p. 39)

Stein, Grover, and Henningsen (1996) found that the thinking and reasoning implied by the task statement are not necessarily the thinking and reasoning students engaged in while working on and talking about the task. Teacher can reduce the cognitive demands

by simply pointing out to students some connections they have seen and letting students memorize them. However, teachers drawing conceptual connections is not one of the factors that support the maintenance of high-level cognitive thinking (Henningsen & Stein, 1997). Rather, teachers can help students understand some mathematical facts as stepping stones and then guide them to construct the connections among mathematical ideas in their own knowledge web. Henningsen and Stein (1997) noted on the importance of scaffolding, based on students' prior knowledge, in helping students to understand and make connections among important ideas:

> Scaffolding occurs when a student cannot work through a task on his or her own, and a teacher or more capable peer provides assistance that enables the student to complete the task alone, but that does not reduce the overall complexity or cognitive demands of the task. Also, teachers can support high-level thinking processes in students by explicitly modeling (or by having a student model) such processes and thinking strategies (Anderson, 1989). (p. 527)

Moreover, there are three other important factors to maintain high-level cognitive thinking demands entailed in a mathematical task: (1) focusing on the thinking processes entailed in reaching the solutions or conclusions rather than the exact answers, (2) giving students enough time to grapple with the important mathematical ideas, and (3) having high expectations and accountability for high-level thinking (Henningsen & Stein, 1997). Martin (2007) gave a similar suggestion as follows:

> Serious mathematical thinking takes time as well as intellectual courage and skills. A learning environment that supports problem solving must provide time for students to puzzle, to be stuck, to try alternative approaches, and to confer with one another and with the teacher. Furthermore, for many worthwhile mathematical tasks, especially those that require reasoning and problem solving, the speed, pace, and quantity of students' work are inappropriate criteria for "doing well". Too often, students have developed the belief that, if they cannot answer a mathematical question almost immediately, then they might as well give up. Teachers must encourage and expect students to persevere when they encounter mathematical challenges and invest the time required to figure things out. In discussions, the teacher must allow time for students to respond to questions and must also expect

> students to give one another time to think without interrupting or showing impatience (NCTM, 2000). (p. 41)

In short, mathematical thinking and learning, from a pedagogical perspective, are highly contextualized. Students learn mathematics much better when tasks are engaging brain exercises instead of time-pressing competitions, when concepts are communicated to them emphasizing the understanding rather than showered on them forcing them to memorize, and when they are able to connect mathematical ideas more pluralistically and tightly than struggling in a chaotic and loosely connected mathematics world.

6.3 The balance between words and ideas

Howe (2001) claimed that "all high school mathematics courses should be seen as loci of reasoning and proof" (p. 45) and "[l]oss of the opportunity for high school students to have an introduction to systematic reasoning would constitute a major institutional failure, but it seems to be happening" (p. 45). King (1992) commented that the certainty of the proved theorems of mathematics holds only due to the complete precision involved in the derivation of them by pure logic, and "truth which comes from deduction and not from observation is possible only by way of complete precision. Pure mathematics is precise or else it is nothing" (p. 60). King (1992) also commented that the totally abstract nature of pure mathematics is often put forward by both mathematicians and non-mathematicians alike as an explanation for the high degree of difficulty the subject presents for all but a small subset of the population, and simultaneously provides some mathematicians "an excuse for their failure to transmit their knowledge of

mathematics to outsiders" (p. 61), who are getting to feel comfortable to believe that "their inability to understand mathematics results from a handicap beyond their control and not from simple failure of nerve or self-discipline" (p. 61). Unfortunately, abstract thinking (nothing more than selective thinking) and abstract objects (including any academic textbooks) are commonplace in our daily life (King, 1992). So it is precision, but not abstraction, that makes mathematics difficult; "precision is unnatural and hard" (King, 1992, p. 62). In King's (1992) view, in mathematics, we can either prove a statement with the given hypotheses or not; there is no third way for us to bluff through. However, Dr. Roy Smith asserted that there is always a third way, namely changing the assumptions, which is standard procedure in any attempt to make progress, though the progress is partial and gradual. Doing mathematics is a research activity. The basic rule of problem solving is to make the problem easier when we are stumped. We do not just set a fixed challenge and refuse to budge over what we want to achieve. Otherwise, we are probably not going to get anywhere. We should not hold the view of our field that it is "all black and white" (personal communication, April 13, 2010). I think this would just be one of the reasons that mathematics enjoys special esteem. Indeed, the desire to be precise prompts us to connect what we have known about mathematics and what we are working on until we get to a solution, a discovery, or even a creation. But the process never ends because there will always be something new to interest us. In other words, mathematics is driven forward by its unsolved problems, and conjectures are a vital part of this process. So I think this agrees with Dr. Smith's opinions and renders that the

precision of pure mathematics noted above by King should not exclude the importance of unsolved problems or unproved conjectures. A good conjecture is extremely valuable and has high status, not as Ernest (1991) considered "informal or unverified conjectures, proof attempts, … to be low status knowledge in mathematics" (p. 99). I would say the creation or discovery of numbers, especially real numbers, is a good illustration of the notion of mathematical precision. In particular, the decimal representation of an irrational number (which can be viewed as an object as well as a process, as I discussed at the end of Chapter 4.2 about the difficulty of understanding a function) makes it less precise than a rational number. For instance, we do not know how to add two infinite decimals. On the other hand, the idea of an irrational number enables us to be more precise than that of a rational number can be in some cases, especially the ones involving the infinite. This might be one difficulty when students deal with real numbers. For example, the aforementioned (in Chapter 3 and Chapter 5.2) notion 0.999…=1 is a difficult one for most students to accept.

In fact, "there is no single theoretical account of mathematics" (Hanna, 1983, p. 63). We should not view mathematics as a static and perfectly–structured system of definitions, axioms, theorems, concepts, and procedures. Though the method of axiomatic system is quite popular, proof, as an essential facet of the axiomatic method, is just one possible way to demonstrate the truth of a statement and the validity of the proof itself within the assumed axiomatic frame of reference, which is not shared by the various schools of mathematical thought. Hanna (1983) proposed that rigorous proof in school mathematics curriculum should be treated as an indispensable topic, a valuable asset of

modern mathematics, instead of a pervasive methodology and mode of presentation. Otherwise, it "may conceal from the student the rich pluralism of modern-day mathematical theory and perpetuate in the student's mind a simplistic and relatively unattractive picture of mathematical practice" (p. 87). Hanna (1983) also argued that rigorous proof plays almost no role in mathematical discovery and creativity, that the proof of a theorem carries little weight in the complex process of the acceptance of the theorem by students or even mathematicians, and that a fully rigorous proof is impossible in some cases due to the inescapable practical limitations. I wrote down the proofs in this thesis mainly for teacher's sake. I, however, believe that understanding important mathematical knowledge must happen prior to bogging down in the bald statement of the rigorous proofs. Teachers may have students focus on those proofs which are not too long but interesting, clever, and involve relatively important mathematical ideas. According to Healy and Hoyles (2000), though it is difficult for the majority of the students to generate valid proofs in domain of algebra on their own, they do value general and explanatory arguments and believe that a valid proof of a statement makes no further work necessary to ascertain the truth of any specific cases of the statement. "The process of proof is undeniably complex, involving a range of student competencies-identifying assumptions, isolating given properties and structures, and organizing logical arguments-each of which is by no means trivial" (p. 396) and certainly analyzing the conclusion guides all of these processes of thinking.

Furthermore, we do not have the right to let mechanical rigor stand in the way of a flexible understanding of a piece of mathematical knowledge. "Rigor and reasoning should always be presented in a way that can be meaningful to the audience. When deep

ideas and reasoning are brought up in service courses, they should be treated in a ‘great ideas’ fashion” (Howe, 2001, p. 45). Instead of going through the proof in a didactic way, teachers may give students an intuitive grasp of those meaningful and powerful ideas behind the abstractions and precisions by a relatively informal narrative exposition so as to help students start with testing and refining their own conjectures, feel the process of proving mathematicians summarized, and “develop more multifaceted competence in proving that includes some deductive reasoning” (Healy & Hoyles, 2000, p. 427). When it comes to mathematics learning and teaching, the above suggestion, I believe, is not in conflict with King’s (1992) proposal about precision in mathematics. This is why I sometimes used a less formal way to state a proof; students deserve to see the essential mathematical thoughts. This suggestion also goes along with the proposal I stated at the beginning of the “algebraic structure” section (Chapter 1.3) that teachers should not let the “formalism” formalize or “format” our students and undermine their understanding, interest, and confidence in mathematics. This is probably why mathematicians seldom give lectures by “reading” a prepared manuscript to the audience; they enjoy a flow of ideas more than a flow of words.

Incidentally, “disproof,” like figuring out why a conclusion is not right or a statement is not true, is another good way to build students’ understanding of mathematical concepts. Martin (2007) proposed that

> When students are allowed to examine and critique incorrect solutions or strategies, counterexamples and logical inconsistencies can naturally surface. This process of analyzing solutions instead of relying on teachers to validate them can enhance students' abilities to think critically from a mathematical perspective. (p. 47)

I like King’s (1992) comment that "the ideas brought forth from the unconscious and handed over to the conscious invariably possess the stamp of mathematical beauty" (p.

139). I will use the two proofs that $\sqrt{2}$ is not rational (given in Chapter 5.3) as an example to illustrate this. We suppose $\sqrt{2}$ is a rational number and let it be in the lowest term. To start with this step we automatically or consciously use the definition of rational numbers and the method of proof by contradiction; this is nice but not as brilliant as the move from the equation $n^2 = 2m^2$ to the contradiction, because it is this equation that motivates the unconsciously further progress of the proof before people know what to do next, and then the pleasure or excitement continues until they consciously reach the exact contradiction. The beauty of the proof by the rational root theorem is perhaps that people stare at the algebraic fact $\sqrt{2} \cdot \sqrt{2} = 2$ unconsciously and finally pull out a more sophisticated proof transferring their work field from the system of integers to the world of polynomials.

Mathematical intuition or unconsciousness is the source of creation or discovery of a theorem; we use the axiomatic method and logic to help us organize and analyze more deeply the consequences of what our intuition has suggested; A proof, discovered by intuition, preceded by substantial mathematical knowledge accumulation and arduous mathematical work or thinking, filled out with conscious confirmation details, is a process that we apply to test those suggestions of our intuition, and is an elaboration to supply communication of any kind.

6.4 Last remarks to share with both students and teachers

Mathematical thinking is more than the process or abilities utilized in absorbing some piece of mathematics or in solving some mathematical problem. Mathematical thinking is closely associated with a consideration of the beauty of mathematics; it can be

an impulse of new instruction or even curricula of mathematics. King (1992) proposed that people "need experience and training in how to look at mathematics just as they need experience and training in how to listen to Beethoven. The natural place for this instruction is in the mathematics classroom" (p. 139). Students should be given opportunities to *see* mathematics as the mathematicians *see* it; mathematical concepts must be motivated for the students as they are motivated for the mathematicians; mathematicians are fallible, so are students (and teachers). For example, students should experience mathematics like research mathematicians using prior knowledge and moderate technology to observe mathematical phenomena, trusting or distrusting their intuitions, trying to solve the problems or make guesses, trying to use their analytical powers to justify those conjectures they made, learning to get used to the abstract versions, and applying the verified conclusions to new, harder problems. Therefore, as King (1992) claimed, it will require our school mathematics teachers to be "people who themselves have been personally touched by mathematics deeply enough to have some chance at communicating to their students a semblance of the excitement of the subject" (p. 143). I like Mortimer Adler's notion of two kinds of beauty: "enjoyable beauty" and "admirable beauty" (as cited in King, 1992, p. 177). It is understandable that mathematics does not have enjoyable beauty to most people just as some people do not think classical music is as enjoyable as rock music. It is in part a matter of personal taste. But most people (including high school students) should be properly influenced to appreciate or at least believe the admirable beauty of mathematics, which has been sensed and judged by experts who "learned enough or experienced enough to have developed superior taste" in mathematics (King, 1992, p. 177). I hope that teachers and teacher educators could work

together to help our students see that mathematics has “accessible beauty,” something I have tried to illustrate in this thesis.

I talked about algebraic structures at the very beginning of this thesis. Whether an object belongs to a structure depends on whether its relations with the substantiated objects of the structure is consistent with the rules of the structure. If an object cannot play within the rules already set for the structure, then the object is not a member of the structure itself. Our society can also be considered as such a "structure." Students know they have to work out a way to fit themselves in the society to survive or even thrive. What if they are taught with the “accessible beauty” of mathematics and motivated to want to treat mathematics (at least school algebra) as a structure to adjust themselves for?

Let us go back to see one more example about some of the basic algebraic rules. Someone says that he can prove 2 = 1 in any sense, providing the following proof:

1) a = b for some a’s and b’s (thinking about modulo arithmetic as an example, or for simplicity, $1=\sqrt{1}$ (which we have shown before); a and b just look different, and they are just two different numerals for two equivalent things),

2) a+a = b+a (doing the same thing to two equivalent things),

3) i.e., 2a = b+a,

4) 2a-2b = b+a-2b (still doing the same thing to two equivalent things),

5) 2(a-b) = a-b (distributivity),

6) 2 = 1 (cancelling a-b on both sides of the equivalence relation),

as desired. Done.

It seems that 2 = 1 has really been proved. But, a = b yields a-b = 0. The proof is a trap. We have been told that we cannot divide by 0. But why? We will appeal to the definition

of division (i.e., we call a number x/y if it is a *unique* number whose product with y is x). Thus to have a number called 2/0, for example, it would have to multiply 0 into 2, which never happens. In other words, if 0 is the defined additive identity, and if we want to have the axiom that a(b+c) = ab+ac, then for every number a, we have a(0) = a(0+0) = a(0)+a(0), so subtracting a(0) from both sides we see that 0 = a(0). Thus, if we want to have subtraction and distributivity of multiplication, then we must have a(0) = 0 for all a. Hence, we can never have any number d such that d(0) equals 2 or whatever non-zero number, i.e., we cannot have any number called 2/0. By the way, this can be seen from the following graph of the function $y = \frac{2}{x}$ with $x, y \in R$. When x approaches 0 in both positive and negative directions, the value of y approaches an uncertainty: infinity or negative infinity? Hence, we cannot divide by 0; similarly, we cannot divide by a zero polynomial (or any additive identity depending on the context).

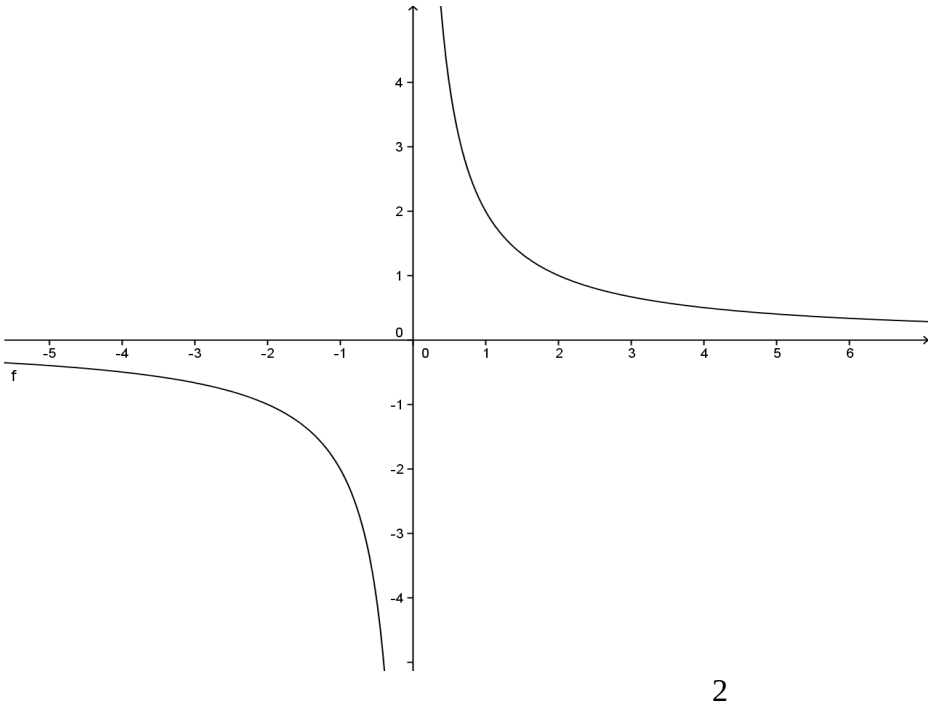


Figure 6.4.1 Partial view of the graph of function $y = \frac{2}{x}$ with $x, y \in R$

On the other hand, to consider the symbol 0/0, we have a different problem: every number qualifies to be called 0/0 because every number multiplies 0 into 0. We do not know what number we are talking about when we write 0/0, which violates the uniqueness part of the definition of division, so we do not want this either. However, in calculus, when we have an expression like x/3x, which gives 0/0 when x = 0, we can make sense of this by canceling the x's first and then we get 1/3.

Now let us try to answer another typical high-school-student question: “Why does a negative times a negative equal a positive?” concerning those basic algebraic rules or axioms. Note that this is essentially a consequent property of an ordered field (such as the real numbers) which we mentioned in Chapter 5.2.2.

Let a and b be two positive elements in an ordered ring. So –a and –b are negative elements in the ring by the closure property of the ring, and ab is positive.

First, we claim b-a = b+(-a). Cancelling b on both sides, we want to show –a = +(-a) as follows. Since 0 is the additive identity, -a = -a+0 = 0+(-a+0) = +(-a).

If we want the distributivity holds, we must have a(0) = a(b-b) = a(b+(-b)) = ab+a(-b). But we know ab-ab = 0. So we must have a(-b) = -ab (1). Similarly (-a)b = -ab (2). And so a(-b) = (-a)b (3).

Let us see 0 = 1-1 = 1+(-1). Since 1 is the multiplicative identity, we have 1+1(-1) = 0, and so 1 = -1(-1), which yields 1 = (-1)(-1) by (2). And (-a)(-b) = (-a)(1)(1)(-b). By (3) we get (-a)(-b) = (a)(-1)(-1)(b). So we have (-a)(-b) = (a)(1)(b) = ab.

Indeed, we just try to agree on what axioms we want to be true and try to live by those rules and to thrive in our world of mathematics (or even in our real life). If you ask a mathematician “What is 1 plus 2?”, he may say: “If we work in the ring Z , and given 1

plus 1 is 2, then 1 plus 2 is 3; if in the ring Z_3 it equals 0." Missing most parts of the chains of deduction in mathematics may oppress or frighten students away from mathematics.

Everyone has taste in mathematics, just like in art or music, simply because taste is having the courage of one's own convictions. The conviction could be that one can do mathematics better than he/she is doing, no matter who he/she is, a high school student or a mathematician, and one can prove it by oneself; the conviction could be that one can always find a good way to help his/her students appreciate mathematics better if he/she wants to prove himself/herself a good teacher.

Middleton and Spanias (1999) proposed that if appropriate practices are consistent over a long period of time, students can and do learn to enjoy and value mathematics. This proposal actually loads most of the student's burden of learning mathematics on to the shoulder of the teacher. It requires the teacher's understanding of fundamental mathematics to be both mathematically profound and pedagogically preferable. As Saul (2008) proposed, in order to implement effective pedagogical techniques to address the difficulties students encounter in learning algebra, which are related to deeper mathematical insights, "teachers need to know more mathematics than they expect their students to know. The question of teachers' content knowledge is an increasingly important one and worthy of our continued attention" (p. 78) (c.f., Ferrini-Mundy & Findell, 2001). On the other hand, as Dr. Roy Smith claimed, no matter how much analyses state that it is crucial to draw connections between different subjects, we cannot teach this way if we give in the pressure to make the classes less demanding (personal

communication, February 9, 2010). It is enlightening to see Dr. Smith's (1997) following proposal

> That student [who has been told something he does not yet understand] is actually receiving instruction not just for the moment, but also for the future; he is being given something to think about which will last him some significant amount of time, and which will repay all the thought he will give to it. … When reading student evaluations of a teacher, how often does one encounter the grateful comment, "He really gave us some provocative questions to think about. I still have not settled them all!"? I would ask, if this comment is missing, can the teacher really be excellent? (p. 13)

I hope that through this thesis the light of mathematical aesthetics of connectivity, logic, precision, efficiency, diligence, unconsciousness, and inspiration will shine brighter in high school mathematics teaching and learning.

REFERENCES

Achieve, Inc. (2005). Rising to the challenge: Are high school graduates prepared for college and work. Washington, DC: Peter D. Hart Research Associates / Public Opinion Strategies. Retrieved from: www.achieve.org/RisingtotheChallenge

Arcavi, A. (2008). Algebra: Purpose and empowerment using concept maps to assess conceptual knowledge. In C. E. Greenes & R. Rubenstein (Eds.), *Algebra and Algebraic Thinking in School Mathematics* (pp. 37-49). Reston, VA: National Council of Teachers of Mathematics.

Ball, D. L. & Bass, H. (2003). Making mathematics reasonable in school. In J. Kilpatrick, W. G. Martin, & D. Schifter (Eds.), *A Research Companion to Principles and Standards for School Mathematics* (pp. 27-44). Reston, VA: National Council of Teachers of Mathematics.

Birkhoff, G. & MacLane, S. (1941). *A survey of modern algebra*. New York: The Macmillan Company.

Boaler, J. & Humphreys, C. (2005). *Connecting mathematical ideas: Middle school video cases to support teaching and learning*. Portsmouth, NH: Heinemann.

Chazan, D. (2000). *Beyond formulas in mathematics and teaching: Dynamics of the high school algebra classroom*. New York: Teachers College Press.

Cuoco, A. A. (1990). *Investigations in algebra: An approach to using Logo*. Cambridge, MA: The MIT Press.

Dweck, C. S. (1986). Motivational processes affecting learning. *American Psychologist*, 41(10), 1040-1048.

Ernest, P. (1991). *The Philosophy of mathematics education*. Bristol, PA: The Falmer Press.

Ferrini-Mundy, J. & Findell B. (2001). The mathematical education of prospective teachers of secondary school mathematics: Old assumptions, new challenges. In *CUPM discussion papers about mathematics and the mathematical sciences in 2010: What should students know?* (pp. 31-41). The Mathematical Association of America.

Foley, G. D. (1998). The role of algebraic structure in the mathematics curriculum of grades 11-14. In *The Nature and Role of Algebra in the K-14 Curriculum: Proceedings of a National Symposium (Washington, DC, May 27-28, 1997)* (pp. 87-88). Washington, DC: National Academy Press.

French, D. (2002). *Teaching and learning algebra*. New York: Continuum.

Gamow, George. (1957). *One, two, three ...infinity: Facts and speculations of science.* New York: The Viking Press.

Hartshorne, Robin. (2000). *Geometry: Euclid and beyond*. New York: Springer-Verlag.

Healy, L. & Hoyles, C. (2000). A study of proof conceptions in algebra. *Journal for Research in Mathematics Education*, 31(4), 396- 428.

Henningsen, M. & Stein, M. K. (1997). Mathematical tasks and student cognition: Classroom-based factors that support and inhibit high-level mathematical thinking and reasoning. *Journal for Research in Mathematics Education*, 28(5), 524 – 549.

Howe, R. (2001). Two critical issues for the math curriculum. In *CUPM discussion papers about mathematics and the mathematical sciences in 2010: What should students know?* (pp. 43-49). The Mathematical Association of America.

Isomorphism (n.d.). In *Wikipedia, the free encyclopedia*. Retrieved from http://en.wikipedia.org/wiki/Isomorphism

Kennedy, J. F. (1960). Remarks of Senator John F. Kennedy at Concord College, Athens, West Virginia, April 27, 1960. Retrieved from http://www.jfklibrary.org/Historical+Resources/Archives/Reference+Desk/Speeches/JFK/JFK+Pre-Pres/1960/002PREPRES12SPEECHES_60APR27B.htm

King, J. P. (1992). *The art of mathematics*. New York: Plenum Press.

Lang, S. (1985a). *Math!: Encounters with high school students*. New York: Springer-Verlag.

Lang, S. (1985b). *The beauty of doing mathematics: Three public dialogues*. New York: Springer-Verlag.

MacGregor, M. & Price, E. (1999). An exploration of aspects of language proficiency and algebra learning. *Journal for Research in Mathematics Education*, 30(4), 449-467.

Martin, T.S. (Ed.). (2007). *Mathematics teaching today: Improving practice, improving student learning*. Reston, VA: National Council of Teachers of Mathematics.

Mathematical Association of America (MAA). (2001). Mathematics and the mathematical sciences in 2010: What should students know? In *CUPM discussion papers about mathematics and the mathematical sciences in 2010: What should students know?* (pp. 1-12). Washington, DC: Author.

Middleton, J. A. & Spanias, P. A. (1999). Motivation for achievement in mathematics: Findings, generalizations, and criticisms of the research. *Journal for Research in Mathematics Education*, 1999, 30(1), 65 – 88.

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: Author.

Rickart, C. (1996). Structuralism and mathematical thinking. In R. J. Sternberg & T. Ben-Zeev (Eds.), *The nature of mathematical thinking* (pp. 285-300). Mahwah, NJ: Lawrence Erlbaum Associates.

Saul, M. (2008). Algebra: The mathematics and the pedagogy. In C. E. Greenes & R. Rubenstein (Eds.), *Algebra and algebraic thinking in school mathematics: 70th YB* (pp. 63-79). Reston, VA: National Council of Teachers of Mathematics.

Shifrin, T. (1996). *Abstract algebra: A geometric approach*. Upper Saddle River, NJ: Prentice Hall.

Smith, R. (1997). On Teaching. *The Mathematics Educator*. 8(1), 12-15.

Stein, M. K., Grover, B. G., & Henningsen, M. (1996). Building student capacity for mathematical thinking and reasoning: An analysis of mathematical tasks used in reform classrooms. *American Educational Research Journal*, 33(2), 455-488.

Tiles, M. (1989). *Philosophy of set theory: An historical introduction to Cantor's paradise*. New York: Basil Blackwell.

Venezia, A., Kirst, M. W.,& Antonio A. L. (2003). Betraying the dream: How disconnected K-12 and postsecondary education systems undermine student aspirations. Stanford University's Bridge Project. Retrieved from www.stanford.edu/group/bridgeproject/betrayingthecollegedream.pdf

Williams, S. E. & Molina, D. (1998). Algebra: What all students can learn. In *The Nature and Role of Algebra in the K-14 Curriculum: Proceedings of a National Symposium (Washington, DC, May 27-28, 1997)* (pp.41-44). Washington, DC: National Academy Press.

Yahoo! Answers (2009). *Questions about abstract algebra?* Retrieved from http://answers.yahoo.com/question/index?qid=20090702124234AAIN73k

APPENDIX

The Cauchy Model of the Real Numbers

The following is the second constructive approach for the real numbers as feedback on my thesis offered by Dr. Roy Smith on Saturday, March 27, 2010, with my annotations in italics in brackets.

Dear Chen,

Here is a pretty little construction of the reals that you may like, if you like Cauchy sequences [*talked about in Chapter 5.3 of this thesis*]. I learned this from the book Modern Algebra by Van der Waerden, English translation of the second German edition, section 67, pages 211-218 [*In later correspondence, Dr. Smith said Waerden might use the model created by Georg Cantor, the man who created set theory, and gave the diagonal argument for the real numbers being more numerous than the rational numbers, which we have talked about in Chapter 5.3*]. It is more abstract than infinite decimals, but that actually makes some things easier to prove. Notice that an infinite decimal is an infinite sequence of finite decimals, with each finite decimal having n decimal places, differing from the next one by less than 1/(10)^n. This is a Cauchy sequence of finite decimals, i.e. of rational numbers. So really a real number is something that can be approximated by rationals, but it takes an infinite number of rationals, and in order to define a unique real number the sequence of rationals should be Cauchy. Of course there

are two sequences that define each finite decimal, one ending in 0's and one ending in 9's. There is no way to choose just one sequence for each real number and this can be confusing. [*In later correspondence, Dr. Smith commented that it is, in fact, possible to pick just one decimal sequence for each real number, but if you do, when you add two numbers there is no guarantee you will get the one you chose as your sum. So there is no way to pick just one sequence so that the sequences you pick will be closed under addition and subtraction and multiplication. That's where the mystery for students comes in. For example, if they use the sequence {1.0, …, 1.0} for 1, and the sequence {0.3, 0.33, 0.333, …, 0.333333…3, … } for 1/3, then when they multiply 1/3 by 3 they get {0.9, 0.99, 0.999, …, 0.999999…9, …} which is not the sequence they chose for 1.*] Thus it is even easier to just use lots of sequences for each real number. [*Though there is no way to write down all Cauchy sequences that define a given number - there are infinitely many, but we can at least try to list some of them to see just how many sequences are possible for each real number. For pi, we could choose say, as first number, any rational number between 3 and 4, then as second number any rational number between 3.1 and 3.2, then as third number any rational number between 3.1 and 3.15, ... , and so on, so the sequence may look like {3.1, 3.14, 3.141, 3.1415, 3.14159, …}. But we could still have a sequence like {2.9, 3.09, 3.139, 3.1409, 3.14149, 3.141589, …}, or {4, 3.2, 3.15, 3.142, 3.1416, 3.14159, …}, or {3.9, 3.19, 3.149, 3.14159, 3.141589, …}, or some other sequences like {1, 3, 3.1, 3.14, 3.141, ...}, {1, 2, 3, 3.1, 3.14, 3.141, …}, {0, 2, 3.4, 3.14, 3.141, ...}, {1, 2, 3, 4, 3, 3.1, 3.14, 3.141, ...}, etc.. We can take any finite sequence whatever and start with it, then switch to the usual sequence, and that will be Cauchy and converge to pi. The monotonicity (increasing or decreasing) of the Cauchy sequence does not matter, and*

especially it does not matter whether the "initial finite sequence" is monotonic or not. But still even these infinitely many possibilities come nowhere near being all possible choices.] We can use this idea to define real numbers as Cauchy sequences of rational numbers, subject to a natural equivalence relation that means the two sequences define the same real number, as follows.

Assume we know about the rational numbers Q and that they form a field. Then define the set C to be the set of all Cauchy sequences of rational numbers. Since the sum and product of two Cauchy sequences is also Cauchy, and the rationals are a ring, we get a ring structure on C by adding and multiplying entries in the sequences. Among the Cauchy sequences in particular are all the convergent sequences of rationals, i.e. rational sequences with a rational limit. Let N = the subset of C of "null sequences", i.e. convergent sequences with limit zero. Then N is an "ideal" of the ring C, i.e. N is a subring but is also closed under multiplication by elements of C, i.e. the product of any element of C with any element of N is contained in N.

Then define the equivalence relation on C where two Cauchy sequences are equivalent if and only if their difference is null. This defines a set of equivalence classes R = C/N, analogous to modular integers Z/nZ. This R is our model of the real numbers.

Then a non null Cauchy sequence has a multiplicative inverse because the absolute values $|a_j|$ of the entries are eventually positive, so the inverse sequence exists after excluding some initial finite sequence of the terms [*for example, to invert (take the*

multiplicative inverse of) the sequence 0, 0, 0, 2, 2, 3, 4.6, …, we use any number such as 1 to substitute initial 0's, and then invert the consequent numbers like inverting 2 to 1/2, and 4.6 to 1/4.6? So the inverse of the sequence is 1, 1, 1, 1/2, 1/2, 1/3, 1/4.6, …].

Moreover if two sequences are not equivalent then their difference sequence is eventually positive or eventually negative so we can define what it means for one equivalence class of sequences to be greater than another, namely they are not equivalent and the elements are eventually greater than the elements of the other sequence.

It is easy to prove that axiom A [*Archimedean property*] holds.

Then one can prove a version of LUB, namely

Theorem: Every strictly monotone increasing (or decreasing) sequence of Cauchy sequences, has a limit.

Proof: sketch. By choosing subsequences one can insure that in each Cauchy sequence all terms after the nth term are closer together than 1/n.

Then one can choose a "diagonal" subsequence by choosing the nth term from the nth sequence, similar to the diagonal construction in Cantor's argument that the reals are uncountable [*again, as shown in Chapter 5.3 in my thesis*].

Then this diagonal subsequence should be a limit of the sequence of Cauchy sequences.

Then one can use axiom A to show rationals are dense in R, and then use this monotone limit theorem to prove the full LUB property [*as shown in Chapter 5.2 in my thesis*].

How do you like that?

Roy