The Major Topics of School Algebra

Wilfried Schmid and H. Wu

June 12, 2008

The following extended discussion of **The Major Topics of School Algebra** was written by us in 2007 for the deliberations of the *Conceptual Knowledge and Skills Task Group* of the National Mathematics Advisory Panel. An abbreviated version now appears in Section V, Sub-section A, of the Task Group's report on Conceptual Knowledge (http://www.ed.gov/about/bdscomm/list/mathpanel/report/conceptual-knowledge.doc). We believe this more elaborate version can still serve to round off the discussion in the report itself.[1]

Symbols and Expressions

- Polynomial expressions
- Rational expressions
- Arithmetic and finite geometric series

Linear Equations

- Real numbers as points on the number line
- Linear equations and their graphs
- Solving problems with linear equations
- Linear inequalities and their graphs
- Graphing and solving systems of simultaneous linear equations

Quadratic Equations

- Factors and factoring of quadratic polynomials with integer coefficients
- Completing the square in quadratic expressions
- Quadratic formula and factoring of general quadratic polynomials
- Using the quadratic formula to solve equations

[1] We are grateful to David Collins for many corrections.

Functions

- Linear functions
- Quadratic functions – word problems involving quadratic functions
- Graphs of quadratic functions and completing the square
- Polynomial functions (including graphs of basic functions)
- Simple nonlinear functions (e.g., square and cube root functions; absolute value; rational functions; step functions)
- Rational exponents, radical expressions, and exponential functions
- Logarithmic functions
- Trigonometric functions
- Fitting simple mathematical models to data

Algebra of Polynomials

- Roots and factorization of polynomials
- Complex numbers and operations
- Fundamental theorem of algebra
- Binomial coefficients (and Pascal's triangle)
- Mathematical induction and the binomial theorem

Combinatorics and Finite Probability

- Combinations and permutations as applications of the binomial theorem and Pascal's Triangle

The preceding list of topics comprises the most basic elements of school algebra. The total amount of time spent on covering these topics would normally be a little more than two years, although how the instruction of these topics is structured throughout high school is a matter to be determined by each curriculum. What is usually called "Algebra I"[2] would *in most cases*, cover the topics in the Symbols and Expressions, Linear Equations, and at least the first two bullets of Quadratic Equations. The usual course called "Algebra II" would cover the rest, although in some cases, the last bullet of Functions (data), the last two bullets of Algebra of Polynomials (binomial coefficients and binomial theorem), and Combinatorics and Finite Probability would be left out. In that case, the latter collection of topics would generally find their way into a course on pre-calculus.

The teaching of algebra, like the teaching of all of school mathematics, must ensure that students are proficient in computational procedures, can reason precisely, and can formulate and solve problems. For this reason, the preceding list of topics should not be regarded a collection of disjointed items neatly packaged to be committed to memory. On the contrary, the teaching should emphasize the connections as well as the logical progression among the topics. The following narrative, written with readers in mind who are already familiar with the curriculum of school algebra, tries to give a brief idea of these connections and the main lines of reasoning underlying them. Because standard texts often treat certain topics incorrectly in the sense of mathematics, a great deal of effort has been spent on detailing what these misconceptions are and how to rectify them.

Symbols and Expressions

It can be argued that the most basic aspect of the learning of algebra is the fluent use of symbols. In this context, the concept of a *variable* occupies a prominent position. In standard algebra texts as well as the mathematics education literature, one rarely finds an explicit definition of what a "variable" is. The absence of a precise definition creates a situation whereby students are asked to understand something which is left largely unexplained, and learning difficulties ensue. Sometimes, a variable is described as a quantity

[2] The standard sequence of "Algebra I", "Geometry", and "Algebra II" is not the only way to organize the high school mathematics curriculum. See, for example, the [Kodaira 1]–[Kodaira 2] series of Japanese texts for a different, but mathematically sound approach.

that changes or varies. The mathematical meaning of the last statement is vague and obscure. At other times it is asserted that students' understanding of this concept should be beyond recognizing that *letters can be used to stand for unknown numbers in equations*, but nothing is said about what it is that students should know "beyond" this recognition. In [NRC2001], for example, one finds a statement that students emerging from elementary school often carry the "perception of letters as representing unknowns but not variables" (p. 270). The difference between "unknowns" and "variables" is unfortunately not clarified. All this adds to the mystery of what a "variable" really is.

In mathematics, a *variable* is an informal abbreviation for "an element in the domain of definition of a function", which is of course a perfectly well-defined concept. If, for example, a function is defined on a set of ordered pairs of numbers, it is referred to as "a function of two variables", and it must be said that, in that case, the emphasis is more on the word "two" than on the word "variables".[3] To the extent that school algebra intends to use the concept of a "variable" beyond this narrow context, and in fact *before* the concept of a function is introduced, we proceed to describe a possible definition of this concept, one that is at least mathematically correct. In the process, we discuss the basic etiquette in the use of symbols, which is after all our main goal.

Let a letter x stand for a number, in the same way that the pronoun "he" stands for a man. Any expression in x is then a number, and all the knowledge accumulated about rational numbers can now be brought to bear on such expressions. In a situation where we have to determine which number x satisfies an equation such as $2x^2 + x - 6 = 0$, the value of the number x would be unknown for the moment and x is then called an **unknown**. In broad outline, this is all there is to it as far as the use of symbols is concerned.

A closer examination of this usage reveals some subtleties, however. Consider first the following three cases of the equality $mn = nm$:

(1) $mn = nm$.

(2) $mn = nm$ for all whole numbers m and n so that $0 \leq m, n \leq 10$.

(3) $mn = nm$ for all real numbers m and n.

The statement (1) *has no meaning*, because we don't know what the symbols m and n stand for. To give an analogy, suppose someone makes the statement, "He is 7 foot 6." Without indicating who "he" refers to, this statement is neither true nor false.[4] It is

[3] In the sciences and engineering, the word "variable" is bandied about with gusto. However, to the extent that mathematics is just a tool rather than the central object of study in such situations, scientists and engineers can afford to be cavalier with mathematical terminology.

[4] It is true if "he" refers to basketball star Yao Ming, but false for Woody Allen.

simply meaningless. If m and n in (1) are real numbers, then (1) is true, but there are other mathematical objects m and n for which (1) is false.[5] On the other hand, (2) is true, but it is a trivial statement because its truth can be checked by successively letting both m and n be the numbers 0, 1, 2, ..., 9, 10, and then computing mn and nm for comparison. The statement (3) is however both true and more profound. As mentioned implicitly above, this is the commutative law of multiplication among real numbers. It is either something you take on faith, or, in some contexts, a not-so-trivial theorem to prove. Thus, despite the fact that all three statements (1)–(3) contain the equality $mn = nm$, they are in fact radically different statements because the specifications for the symbols m and n are different. Therefore a basic rule concerning the use of symbols is that the specifications for the symbols are every bit as important as the symbolic expressions themselves.

Next, consider the solution of the linear equation $3x + 7 = 5$. The usual procedure for solving such equations yields $3x = 5 - 7$, and therefore

$$x = \frac{5-7}{3}$$

There is a reason why we do not write the solution as $\frac{-2}{3}$, because we can also consider $3x + \frac{1}{2} = 13$ and get

$$x = \frac{13 - \frac{1}{2}}{3}$$

Or consider $3x + 25 = 4.6$ and get

$$x = \frac{4.6 - 25}{3}$$

Or consider $5x + 25 = 4.6$ and get

$$x = \frac{4.6 - 25}{5}$$

And so on. There is an unmistakable pattern here: no matter what the numbers a, b, and c may be, the solution of the linear equation $ax + b = c$, with a, b, c ($a \neq 0$) understood to be *three fixed numbers* throughout this discussion, is

$$x = \frac{c-b}{a}$$

We have now witnessed the fact that in some symbolic expressions, the symbols stand for elements in an infinite[6] set of numbers, e.g., the statement that $mn = nm$ for *all*

[5] For example, certain 2×2 matrices.

[6] As mentioned at the beginning of this article, a variable is an element in the domain of a function, and the domain can be finite or infinite. But for school algebra, where functions are those defined on intervals of the number line, saying that a domain is "infinite" suffices for the purpose at hand.

real numbers m and n, while in others, the symbols stand for fixed values throughout the discussion, e.g, the numbers a, b, and c in the linear equation $ax + b = c$. In the former case, the symbols are called **variables**, and in the latter case, **constants**. The main message is, therefore, that

in any symbolic expression, one must specify
precisely what each symbol stands for.

We see that a variable so defined *does not vary or change*. It is simply an element in an infinite set of numbers.

In view of the preceding discussion, we should point out to students in elementary school that the usual statements for the associative laws and commutative laws of addition and multiplication, as well as the distributive law, are examples of the use of variables *in the sense just described* (rather than as quantities that change or vary), e.g.,

$$a + (b + c) = (a + b) + c$$

for all numbers a, b, c. Students' success in algebra would be helped by a natural and gradual acclimatization to the use of symbols before actually taking algebra.

To summarize, students need not be told the historical background of the concept of a variable, but they certainly need to know the italicized message above. Above all, they must be cleared of the misconception that a variable is "a quantity that changes or varies."

Introductory algebra should address the care that must be exercised in handling symbolic expressions. Suppose we are given an expression in a variable x, which we may assume to be *any* number. Because all we know about x is that it is a number but not its exact value, computations with the symbolic expression must be done using only the rules we know to be true for *all* numbers, namely, the associative, commutative, and distributive laws. Doing computations not with concrete numbers such as 5 or 17 or 82, but with an arbitrary number brings into focus the concept of **generality**. It requires that we concentrate on properties of all numbers *in general*. For this reason, beginning algebra is just *generalized arithmetic*. What is important to the learning of algebra is that, while the generality is important, one should not forget that one is dealing with numbers, pure and simple. Since the only numbers students get to know before algebra are the rational numbers, *this underscores the importance of rational numbers for the learning of algebra* (compare [Wu]).

An expression of the following type,

$$a_4x^4 + a_3x^3 + a_2x^2 + a_1x + a_0$$

where x is an arbitrary number and a_4, a_3, $\ldots$, a_0 are constants, is called a **polynomial in *x***, which may be denoted by $p(x)$. Thus for each number x, $p(x)$ is also a number. If $p(x)$, $q(x)$ are polynomials, their quotient $\frac{p(x)}{q(x)}$ is called a **rational expression.** Notice that in the case of a rational expression $\frac{p(x)}{q(x)}$, the number x has to be a number so that $q(x) \neq 0$ to avoid division by 0. Let this be understood in the following discussion. Consider the following sum of two rational expressions:

$$\frac{x^2}{(3x^4+x+2)} + \frac{6}{(x^2+5)}$$

Since each of

$$x^2, \quad 3x^4+x+2, \quad 6, \quad \text{and} \quad x^2+5$$

is a number, the preceding sum can be added as numbers. This is because if we think of x^2 as a number a, $3x^4+x+2$ as a number b, 6 as a number c, and x^2+5 as a number d, then the sum becomes just $\frac{a}{b}+\frac{c}{d}$. The usual addition formula

$$\frac{a}{b}+\frac{c}{d} = \frac{ad+cb}{bd}$$

then leads immediately to

$$\frac{x^2}{(3x^4+x+2)} + \frac{6}{(x^2+5)} \quad = \quad \frac{x^2(x^2+5)+6(3x^4+x+2)}{(3x^4+x+2)(x^2+5)}$$

Two remarks about the preceding paragraph should be made. The first is that the above addition points to the need to pay special attention to the symbolic manipulations of *quotients* of rational numbers such as

$$\frac{\frac{3}{11}+\frac{22}{9}}{\frac{2}{3}\times\frac{7}{5}}$$

These quotients come up naturally, as we have just seen, but the need to address their arithmetic operations has not yet been universally recognized. For example, if $x = \frac{2}{3}$ in the preceding sum of rational expressions, then we have

$$\frac{\frac{4}{9}}{(3\times\frac{16}{81}+\frac{2}{3}+2)} + \frac{6}{(\frac{4}{9}+5)} \quad = \quad \frac{\frac{4}{9}(\frac{4}{9}+5)+6(3\times\frac{16}{81}+\frac{2}{3}+2)}{(3\times\frac{16}{81}+\frac{2}{3}+2)(\frac{4}{9}+5)}$$

Students should be entirely at ease with the computations of such quotients. The school curriculum prior to algebra should discuss these computations with care.

The second remark is that we have made repeated references to the variable x as a number, and it is time that we clarify what a number is. In mathematics, a number is a point on the **number line**, usually taken to be a horizontal line, in which the integers are positioned as equi-spaced points with 0 to the right of -1, 1 to the right of 0, 2 to the right of 1, etc. (One may take the number line to be the ordinary x-axis in the coordinate plane.) In the mathematics literature, the number line is called **the real line**, and a number is called a **real number**. The number line should be a central topic in the school mathematics curriculum, especially in the grades leading up to algebra. On the other hand, certain numbers, namely, the **rational numbers**, are treated more thoroughly than others in school mathematics. Rational numbers are located on the number line in the following way. Let m, n be *positive* integers. Divide every segment between consecutive integers into n segments of equal length, so that all the division points now form a new sequence of equi-spaced points. The m-th division point to the right of 0 is $\frac{m}{n}$, while the m-th division point to the left of 0 is the negative rational number $-\frac{m}{n}$. As m and n take on all possible values among positive integers, we get all the nonzero rational numbers. **Irrational numbers**, which are the points on the number line not among the points $\{\pm\frac{m}{n}\}$ above for arbitrary positive integers m and n, are basically no more than a name in the school mathematics curriculum. Their arithmetic, such as the meaning of $\frac{\pi}{\sqrt{2}}$, is taken on faith, as we proceed to explain.

It is unfortunate, but true nevertheless, that by tradition, school mathematics does not make explicit its restriction to only rational numbers in mathematical discussions. The previous example of the addition of two rational expressions, i.e.,

$$\frac{x^2}{(3x^4+x+2)} + \frac{6}{(x^2+5)} \quad = \quad \frac{x^2(x^2+5)+6(3x^4+x+2)}{(3x^4+x+2)(x^2+5)}$$

serves to illustrate this point. Indeed, since it is supposed to hold for *all* numbers, we may let x be π and the equality would still be valid. But of course, school mathematics cannot give meaning to the numbers

$$\frac{\pi^2}{3\pi^4+\pi+2} \quad \text{and} \quad \frac{6}{\pi^2+5},$$

much less discuss how to add them. This should not be interpreted as a fault of school mathematics because any serious discussion of irrational numbers would be unsuitable for K–12. Once we know how to add these two quotients when x is a rational number, then their addition when x is irrational is justified in advanced mathematics by considerations of the "extension of continuous functions from rational numbers to real numbers". While we should leave such considerations out of school mathematics, it is good education nonetheless to make explicit this extrapolation from rational numbers to all real numbers. This can be done without undue effort by appealing to what we call the **Fundamental Assumption of School Mathematics** (**FASM**):

> All the information about the arithmetic operations on fractions can be extrapolated to all real numbers.

Using FASM, we see how to perform the preceding sum of rational expressions in x: we simply pretend that x is a rational number, add, and then have the assurance that the result would be valid for *all* numbers x. One should keep FASM in mind in each encounter with the real numbers in school mathematics.

It is in this setting of learning how to use symbols that we can discuss the many common **identities**, i.e., equalities of symbolic expressions which are true for all numbers. For example, an application of the distributive law for numbers yields:

$$(x-y)(x+y) = x^2 - y^2 \quad \text{for all numbers } x \text{ and } y$$

Written backwards, this identity becomes

$$x^2 - y^2 = (x-y)(x+y) \quad \text{for all numbers } x \text{ and } y$$

In this form, this identity is said to give a **factorization** of the expression $x^2 - y^2$. Similarly, the distributive law gives the fairly obvious fact that

$$(2x+15)(5x-12) = 10x^2 + 51x - 180 \quad \text{for all numbers } x$$

Now we write it backwards and it looks much less obvious:

$$10x^2 + 51x - 180 = (2x+15)(5x-12)$$

We have just factored $10x^2 + 51x - 180$. In an algebra classroom, the factoring of quadratic polynomials,[7] such as the preceding example, is an interesting exercise at least for quadratic polynomials with simple integer coefficients. It is pointless, however, to elevate it to be a major topic in beginning algebra, as sometimes happens in school classrooms. If a little is good, a lot is not necessarily better. We will have more to say on this topic once we come to quadratic equations.

Using the distributive law and nothing else, we can generalize $x^2 - y^2 = (x-y)(x+y)$ to:

$$\begin{aligned} x^{n+1} - y^{n+1} &= (x-y)(x^n + x^{n-1}y + x^{n-2}y^2 + x^{n-3}y^3 + \cdots + xy^{n-1} + y^n) \\ &\quad \text{for } \textit{any} \text{ two numbers } x \text{ and } y, \text{ and } \textit{any} \text{ positive integer } n. \end{aligned}$$

[7] In school mathematics, it is customary to refer to quadratic polynomials as "trinomials". We avoid using this terminology here for two reasons. One is that in the context of the completely accepted terminology of the *binomial theorem*, "trinomial" is confusing. The other reason is that the term "trinomial" is not used in higher mathematics.

This is an important identity, and it belongs to introductory algebra because its proof is just a routine expansion using the distributive law. (Students generally need plenty of exercises on the distributive law.) For example, letting $y = 1$, we get another identity in x alone:

$$x^{n+1} - 1 = (x-1)(x^n + x^{n-1} + x^{n-2} + \cdots + x^2 + x + 1)$$

for all numbers x and for all positive integers n. If $x \neq 1$, multiplying both sides by the number $\frac{1}{x-1}$ and switching the left and the right sides give:

$$1 + x + x^2 + x^3 + \cdots + x^n = \frac{x^{n+1} - 1}{x - 1}$$

for any number $x \neq 1$, and for any positive integer n. This is of course the formula of the **finite geometric series**.

This summation formula is usually taken up near the end of the study of algebra in high school, whereas we have seen that, as an exercise in the use of symbols, it belongs to the very beginning of algebra. There is no reason for the delay, all the more so because this formula is important in so many areas of pure and applied mathematics.

Parallel to the consideration of the geometric series is the **arithmetic series**, which is the sum of the first n terms of a sequence of "equi-spaced" numbers, i.e., a, $a + b$, $a + 2b$, $a + 3b$, etc., where a and b are fixed numbers. The arithmetic series also serves as a good illustration of the use of symbols. If we sum the first n terms of a, $a + b$, $a + 2b$, $a + 3b$, . . . , the associative and commutative laws imply that

$$a + (a + b) + \cdots + (a + (n-1)b) = na + \{1 + 2 + \cdots + (n-1)\}b$$

But *twice* the sum $1 + 2 + \cdots + (n-1)$ equals

$$\begin{aligned} 2\{1 + 2 + \cdots + (n-1)\} &= \{1 + 2 + \cdots + (n-1)\} \; + \; \{(n-1) + \cdots + 2 + 1\} \\ &= (1 + (n-1)) + (2 + (n-2)) + \cdots + ((n-1) + 1) \\ &= n + n + \cdots + n \quad ((n-1) \text{ times}) \\ &= (n-1)n \end{aligned}$$

Hence, we have the formula of the arithmetic series:

$$a + (a + b) + \cdots + (a + (n-1)b) = na + \frac{1}{2}(n-1)nb$$

Linear Equations

The foundational skill in the study of linear equations is the solution of so-called **linear equations of one variable**, which are any equalities of the form (or can be brought to this form by the associative, commutative, and distributive laws) $ax+b = cx+d$, where a, b, c. d are given constants and x is the variable. The equation implicitly carries with it the question: *which numbers x would make the equality valid?* Note that we are dealing with numbers, period. It is common in school mathematics to classify these equations as "one-step" equations, "two-step" equations, and "multi-step" equations. Such a classification, and therewith this particular approach to linear equations, is misleading on many levels. It leads to a fragmented conception of a simple object (linear equation of one variable). It blurs the main idea behind the solution. It also breeds the misconception that the simple symbolic manipulations needed for the solution are an end in itself. Education research in algebra has even given such manipulations a name, *transformational activities.* In teaching the solution of linear equations, we should give students a better mathematical perspective, and it goes as follows.

The correct solution method begins with an assumption: ***suppose there is a solution $\boldsymbol{x_0}$***. Then this number x_0 satisfies $ax_0 + b = cx_0 + d$, and we emphasize once again that this is an equality between two numbers, $ax_0 + b$ and $cx_0 + d$, no more and no less. We want to find out what x_0 is. The basic observation is that if the equality happens to be $ax_0 = d$ (i.e., $b = c = 0$) and $a \neq 0$, then there is no doubt about what x_0 must be: multiplying both sides by $\frac{1}{a}$, we get $x_0 = \frac{d}{a}$. Thus if x_0 *is isolated on one side,* the equation can be easily solved. In general, the equality can be brought to this form anyway by adding $-cx_0 - b$ to both sides to get $ax_0 - cx_0 = d - b$, so that $(a - c)x_0 = d - b$. Therefore, assuming that $a - c \neq 0$, we get $x_0 = \frac{d-b}{a-c}$ as before.

So far, what we have achieved is just that if there is a solution x_0, then it must be $\frac{d-b}{a-c}$. *We do not know that it is a solution, yet.* Of course it is now a simple matter to directly verify that it is a solution, i.e.,

$$a\left(\frac{d-b}{a-c}\right) + b = c\left(\frac{d-b}{a-c}\right) + d$$

With this general assurance, we now know that the usual method of solution as described in textbooks, which directly computes with the variable x, is indeed correct when properly interpreted. But we must not forget the truth behind these symbolic computations with x, the fact that we are actually computing with a fixed number x_0 which is the presumed solution.

A noteworthy feature of the preceding solution is that the computations using x_0 had to use the associative, commutative, and distributive laws on a number (i.e., x_0) whose value was not known at the time. For example, we had $ax_0 - cx_0 = (a - c)x_0$ *without* knowing anything about x_0 beyond the fact that it is a number. This is one of the reasons we want these laws to be valid for *all* numbers (cf. the above discussion of generality). This is also a very good way to convince students why these laws are important, and are not merely facts to memorize for standardized tests.

A major topic in beginning algebra is the relationship between a **linear equation in two variables** x **and** y, $ax + by = c$, and its graph. We call attention to the many gaps in the usual mathematical discussion of these equations and associated linear inequalities. The first is the lack of a correct definition of the **slope** of a line L — here as in the following, we use "line" as a shorthand for "straight line". It needs to be shown that the slope of L defined by two chosen points P and Q on L is in fact independent of the choice of P and Q. In this case, it is not merely mathematical correctness for its own sake. Indeed, knowing this independence leads to the awareness that, in each situation, we are free to choose the two points that suit us best for the purpose of computing the slope. Sometimes, being able to make such a choice is the difference between success and failure.

The proof of this independence is based on the so-called *AAA criterion for the similarity of triangles*, to the effect that two triangles are similar if their angles are pairwise equal. This is the reason that students must be taught the basic facts about similarity before coming to algebra. It is not necessary for students to know how to prove the AAA criterion before taking up algebra. On the contrary, it suffices that they learn how to use it fluently for now, and wait for an explanation in a later course in geometry. The learning of mathematics need not be strictly linear.

A second gap is the precise definition of the **graph of** $\boldsymbol{ax + by = c}$ as the set of *all* the points (x', y') whose coordinates satisfy the equation, i.e., $ax' + by' = c$. Students generally know this definition by rote, but not why this definition is critical to the proof of the basic theorem of linear equations in two variables: *the graph of* $ax + by = c$ *is a line, and any line is the graph of some linear equation of two variables.* Another critical component of the proof is again to know when two triangles are similar. The importance of this proof stems from the fact that it explains the genesis of the many forms of the equation of a line (point-slope form, slope-intercept form, etc.) that satisfies some prescribed conditions, e.g., passing through two given points, or passing through a given point with a prescribed slope. Knowing this proof makes it unnecessary to memorize these different forms by brute force. At the moment, the relationship between a linear equation and its graph remains a black box to many students, which then makes brute force memorization the only recourse for learning this topic.

Associated with linear equations in two variables are linear inequalities. Again, one must first give a precise definition of the **graph of an inequality** $ax + by \leq c$ as well as a precise definition of a **half-plane** of a line before proving that such a graph is a half-plane of the line defined by the associated equation $ax + by = c$. There are many ways to handle this theorem. A drastic way to cut through the subtleties is to simply define a half-plane to be the graph of a linear inequality and then give many examples and ample discussions to make this drastic step reasonable. A more mundane way is to define the two half-planes of a non-vertical line to be the part of the plane above the line and the part below it. (The case of a vertical line is easily handled: its two half-planes are the points to its left and to its right.) Then one can give a simple proof of the preceding theorem. Once again, we emphasize that students need to learn precise definitions of the key concepts *and* how to make use of these definitions.

It now appears to be standard practice to take up linear programming as an application of linear inequalities. The fact (which can be presented without a formal definition of the concept of a function) that a linear function of two variables $f(x, y) = ax + by + d$ assumes it maximum and minimum in a convex polygonal region $\mathcal{P}$ at a vertex can now be explained using the fact that if f assumes a maximum in $\mathcal{P}$ at a point (x_0, y_0), say $f(x_0, y_0) = h$, then $\mathcal{P}$ must lie completely in a half-plane of the line defined by $f(x, y) = h$, i.e., $ax + by = h - d$. The proof of this fact depends squarely on the theorem of the preceding paragraph, namely, that the graph of a linear inequality is a half-plane.

The subject of simultaneous linear equations (to be called **linear systems**) can be approached in a very straightforward manner if the interplay between the algebra of a linear equation and the geometry of its graph is explained clearly and carefully. This would require, for example, that the meaning of the **solution** of a linear system to be *made explicit*, so that the reason why the (coordinates of the) point of intersection of the lines defined by the individual equations provide a solution is clearly articulated. Again, it is a matter of knowing the precise definition of a *solution* and how to use this definition to prove theorems. This knowledge would allow students to understand why one can use graphing calculators to get the solution of a linear system. Many students at present do not have access to this knowledge.

There is a misconception among students that the usual method of solution of a linear system **by substitution** (or equivalently, by the elimination of a variable) is an exercise in the symbolic manipulation of variables. This requires a comment, which is not unlike the one made earlier in connection with the solution of a linear equation of one variable. When done *correctly*, the method of substitution is strictly a computation with fixed numbers, no more and no less. What the substitution method does is not to produce a solution, but rather to show that ***if there is a solution*** $(\boldsymbol{x_0}, \boldsymbol{y_0})$, what the values of x_0 and y_0 must be. The usual symbolic computation, if presented correctly, is done with the fixed numbers x_0 and y_0 (rather than the variables x and y). Once the explicit values of x_0 and y_0 have been found by the usual method, then they can be substituted into the original linear system to show that they are indeed solutions. Again, when done carefully, the computation shows clearly why the commutative, associative, and distributive laws are important. This harks back to the concept of generality mentioned earlier.

The relationship between solving linear systems and obtaining the intersection of lines is an important one in school algebra. The theorem that two lines are parallel if and only if they have the same slope should be clearly explained. Once we know when two lines do or do not intersect depending on whether they have the same slope or not, this information then gives precise conditions for the solvability of a linear system. This interplay between algebra and geometry is a major unifying theme in mathematics at all levels, and deserves to be emphasized in school algebra.

The availability of the technique of solving linear systems introduces a whole new class of word problems to algebra. Many word problems in grades six and seven, such as the coin problems or problems involving chickens and rabbits in a cage, can be done more transparently by the use of linear systems.

Quadratic Equations

Before approaching quadratic equations, students need some firm grounding in the concept of a square root, which is more subtle than usually realized. Given a positive number s, then there is one and only one *positive* number r so that $r^2 = s$. The fact that there *is* such an r is not trivial to prove, and, in fact, cannot be proved in school mathematics.[8] But the fact that there is at most one such r, i.e., the *uniqueness* of this r, can be demonstrated with care to eighth graders and beyond. By definition, this r is called **the square root of $\boldsymbol{s}$**, and is denoted by $\boldsymbol{\sqrt{s}}$. Thus by the definition of the notation, $\sqrt{s}$ *is always* ≥ 0. From the uniqueness of the square root, one concludes the critical fact that

$$\sqrt{ab} = \sqrt{a}\sqrt{b} \quad \text{for all positive } a,\, b$$

[8] It is possible, however, to convey the main idea through the use of the number line.

This fact is usually either left unexplained, or relegated to verification by calculator for a few special cases. We recommend that it be carefully explained (proved), not only because students can learn from such reasoning, but also that they *need to learn* such reasoning to prepare them for advanced mathematics.

Now let x be a number. Then a **quadratic equation in one variable** is an equality of the form (or can be brought to this form by the associative, commutative, and distributive laws) $ax^2 + bx + c = 0$ which asks for all the real numbers x that make this equality valid (a, b, c are constants and $a \neq 0$). A **solution**, or a **root**, of the equation is a real number x_0 which **satisfies** the equation, i.e., $ax_0^2 + bx_0 + c = 0$. To **solve the equation** is to find all the solutions of the equation. It may be pointed out to students at the outset that, unlike the case of a linear equation in one variable, some quadratic equations do not have solutions, e.g., $x^2 + 3 = 0$.

The method of solution dates back to the Babylonians some four thousand years ago. It consists of two steps:

> (A) One can solve all quadratic equations of the form $a(x + p)^2 + q = 0$, if it has a solution.
>
> (B) All quadratic equations can be brought to the form in (A) by the use of the associative, commutative, and distributive laws.

Solving (A) is relatively straightforward: if it has a solution, it has to be one of two possibilities:

$$r_1,\ r_2 = -p \pm \sqrt{-\frac{q}{a}}$$

Again, the precise method of arriving at these solutions is worth repeating. First, ***if we assume that there is a solution $\boldsymbol{x_0}$***, then we have an equality among numbers: $a(x_0+p)^2+q = 0$, from which we conclude that x_0 must have one of the two possible values as above. This does *not* say that these values are indeed solutions of $a(x + p)^2 + q = 0$. For that, we have to directly verify that

$$a(\left(-p \pm \sqrt{-\frac{q}{a}}\right) + p)^2 + q = 0$$

This is a routine computation. The point, however, is that solving a quadratic equation involves nothing but computation with fixed *numbers*. The consideration of "variables" does not enter.

A little reflection is in order at this point. For $-p \pm \sqrt{-\frac{q}{a}}$ to be solutions (i.e., points on the number line), necessarily $q/a \leq 0$, because there is no square root of a negative number in the sense defined above. Conversely, as we have seen, if we know $q/a < 0$, then $-p \pm \sqrt{-\frac{q}{a}}$ *are* solutions. Thus we have proved, incidentally, that

$$a(x+p)^2+q=0 \text{ has solutions } \iff q/a \le 0.$$

The proof of (B) of course depends on the technique of **completing the square**. On the one hand, this step is decisive for the solution, and is not one that yields easily to discovery learning. On the other hand, once the idea of completing the square is accepted, its implementation is quite routine. Indeed, if we know in advance that ax^2+bx+c can be brought to the form $a(x+p)^2+q$ for some suitable p and q, then expanding the latter by the distributive law gives

$$ax^2+bx+c = ax^2+2apx+(ap^2+q)$$

Comparing the coefficients of both sides, we conclude that letting

$$p=\frac{b}{2a} \quad \text{and} \quad q=c-\frac{b^2}{4a}$$

should work. Indeed it does, as a simple computation shows. Another way, equally valid, of looking at this process is to ask what can be put in the box to make the expression within the parentheses a square:

$$a(x+p)^2+q = a(x^2+2px+\boxed{}) - a\boxed{} + q$$

Bearing in mind that $(x^2+2kx+k^2)=(x+k)^2$, we easily come to the same answer. Recalling the earlier formula for the roots of $a(x+p)^2+q=0$ as $-p \pm \sqrt{-\frac{q}{a}}$, we have now obtained the famous **quadratic formula** for the roots of $ax^2+bx+c=0$:

$$r_1,\ r_2 = \frac{-b\pm\sqrt{b^2-4ac}}{2a}$$

(Note that this derivation requires the use of the identity $\sqrt{AB}=\sqrt{A}\sqrt{B}$.) The earlier reasoning about the solvability of $a(x+p)^2+q=0$ leads to a similar conclusion:

$$ax^2+bx+c=0 \text{ has solutions } \iff b^2-4ac \ge 0.$$

Completing the square is a basic technique in school mathematics, and its significance goes beyond getting the roots of a quadratic equation. When we get to the discussion of quadratic functions, for example, we will see how it leads to a complete understanding of their graphs.

It remains to bring closure to this discussion by mentioning that, if the quadratic polynomial ax^2+bx+c $(a\neq 0)$ already comes in the form $a(x-r_1)(x-r_2)$ for some numbers r_1, r_2, then solving the quadratic equation $a(x-r_1)(x-r_2)=0$ becomes extremely pleasant: the roots are r_1 and r_2. This is because the following basic fact about numbers implies that $x-r_1=0$ or $x-r_2=0$.

If $pq = 0$ for two numbers p and q, then one of p and q is 0.

The proof of this fact should be emphasized in the classroom: By FASM, we may act as if p and q are both rational numbers, which may be taken to be quotients of integers. So suppose $p \neq 0$, we will prove that $q = 0$. Indeed, multiply both sides of $pq = 0$ by the reciprocal $\frac{1}{p}$ of p, we get immediately $q = 0$, as desired.

The long and short of it is that, if a quadratic polynomial is already factored as a product of linear polynomials, then the corresponding quadratic equation can be solved by inspection. This is certainly one reason why the factoring of quadratic polynomials is of interest, and this fact can be used to motivate the usual exercises on factoring such polynomials.

It is a remarkable fact that, conversely, if we can solve a quadratic equation, then we also obtain a factorization of the corresponding quadratic polynomial. Let $ax^2+bx+c = 0$ be given. Denoting the roots given by the quadratic formula by r_1 and r_2 as above, then we claim that the following *identity* in x is valid:

$$ax^2 + bx + c = a(x - r_1)(x - r_2) \quad \text{for all } x$$

Why this is remarkable is that it allows us to recover the whole expression $ax^2 + bx + c$ completely as the product of $a(x-r_1)(x-r_2)$ as soon as we get to know *the two values* of x at which ax^2+bx+c becomes 0 (i.e., $ar_1^2+br_1+c = 0$ and $ar_2^2+br_2+c = 0$). This can be explained as follows. From the explicit expression of the roots r_1, r_2 of $ax^2+bx+c = 0$ in terms of a, b, and c (i.e., the quadratic formula), we obtain the following interesting relations between the roots and the coefficients of a quadratic polynomial:

$$r_1 + r_2 = -\frac{b}{a} \quad \text{and} \quad r_1 r_2 = \frac{c}{a}$$

Therefore,

$$\begin{aligned} a(x - r_1)(x - r_2) &= a(x^2 - (r_1 + r_2)x + r_1 r_2) \\ &= a(x^2 + \frac{b}{a}x + \frac{c}{a}) \\ &= ax^2 + bx + c \end{aligned}$$

which is the asserted identity.

In particular, we see that factoring quadratic polynomials can be made entirely mechanical: just use the quadratic formula to get the roots and apply the preceding identity. Students should be made aware of this perspective to the factoring of quadratic polynomials.

Many students do not see that the identity $ax^2 + bx + c = a(x - r_1)(x - r_2)$ requires any proof. This is an example of a not-uncommon confusion between a simple statement (*if* $ax^2 + bx + c = a(x - r_1)(x - r_2)$ *for all* x*, then* r_1 *and* r_2 *are the roots*) and its not-so-simple converse (*if* r_1 *and* r_2 *are the roots, then* $ax^2 + bx + c = a(x - r_1)(x - r_2)$ *for all* x). From this instance, we can see the need to stress reasoning in school mathematics.

If the quadratic polynomial is $x^2 + bx + c$, i.e., if $a = 1$, then the preceding relations between the roots and the coefficients simplify to

$$r_1 + r_2 = -b \quad \text{and} \quad r_1 r_2 = c$$

These attractive relations between roots and coefficients have generalizations to polynomials of any degree.

The ability to solve quadratic equations greatly enlarges the range of word problems, which up to this point involve only linear phenomena. One particular example is the problem of finding the rectangle having the greatest area among all rectangles with a fixed perimeter.

Functions

The next major topic in algebra is the concept of a function, its definition and the detailed study of linear and quadratic functions. Students usually have some trouble coming to grips with this concept. Other than the need of many examples to illustrate its many-sided ramifications, students should also be shown why functions are indispensable. One can ask how to describe the temperature of a cup of freshly brewed coffee in its first 15 minutes, for example, or the position of a piece of chalk thrown across the classroom. Much of mathematics grows out of necessity, and students should get to know this fact.

Two common practices in the teaching of functions should be avoided. One is to say that a function can be represented by (pictures of) graphs, tables, rules and formulas. Since in school algebra the functions are almost always real-valued functions defined on the number line, tables and the pictures of graphs can *suggest* certain aspects of a function, but can *never* give a complete representation of a function. A second one is to over-emphasize the importance of the concept of a *relation*, especially the difference between a relation and a function. Some books even go so far as to define the *domain* and *range* of a relation. From the point of view of what mathematics needs, the concept of a relation deserves to be mentioned for students' benefit, but it is far from a main topic in school algebra. We repeat what was said earlier about factoring quadratic polynomials: if a little bit is good, it does not follow that a lot is better.

It may seem repetitious, but it needs to be said once more that one must give an explicit definition of the **graph** of a function and use it to prove theorems. For a real-valued function of one variable, its graph is the subset of the plane consisting of all ordered pairs $\{(x, f(x))\}$ where x is a member of the domain of f.[9] Emphasizing the definition of a graph addresses a frequently-asked-question by students: if the graph of f crosses the x-axis at $(m, 0)$, why is m a solution of $f(x) = 0$? This is because $(m, 0)$, being a point on the graph of f, must be of the form $(t, f(t))$, by the definition of the graph. Therefore $(m, 0)$ is equal to $(m, f(m))$, so that $0 = f(m)$, and m is a solution of $f(x) = 0$.

The graphs of **linear functions**, that is, those of the form $f(x) = cx + k$ (c, k being constants), are lines. This follows from what we know about the graphs of linear equations in two variables, because the graph of a linear function $f(x) = ax + k$ is the same as the graph of the linear equation of two variables $y = ax + k$. A special class of linear functions, those **without constant term** (i.e., k), are especially important in middle school mathematics. They underlie all considerations of *constant rate*: constant speed, for example, is the statement that there is a constant v, so that if the distance traveled from the origin at time 0 to time t is $f(t)$, then $f(t) = vt$. It also underlies all the problems connected with **proportional reasoning**, but since this topic is not well understood in mathematics education at the moment, we pause to give a brief discussion.

The general understanding of proportional reasoning (cf. [NRC2001], p. 241–244), to the effect that it is about "understanding the underlying relationships in a proportional situation and working with these relationships" (*loc. cit.*, p. 241), appears to be related to some misconception about the formulation of mathematical problems. To explain this misconception, consider a prototypical proportional-reasoning problem such as the following.

> A group of 8 people are going camping for three days and need to carry their own water. They read in a guide book that 12.5 liters are needed for a party of 5 persons for 1 day. How much water should they carry?

On one level, proportional reasoning in the way it is usually understood is about the ability to infer from the given data that, the amount of water consumed by n individuals

[9] Of course, what we call the *graph* of f is usually taken in advanced mathematics to be *the definition of the function itself*, i.e., a function is just a particular collection of ordered pairs of real numbers. In the New Math period, such a definition of a function was introduced into the school curriculum. The fact that this decision was ill-advised is now commonly accepted as such.

per day is proportional to n. In symbolic terms, if $f(n)$ is the amount of water consumed by n individuals per day (where n is a positive integer), then a student who is capable of proportional reasoning would supposedly conclude that

$$\frac{f(m)}{m} = \frac{f(n)}{n}$$

for all positive integers m and n. *This is an unreasonable expectation, because there is no logical reasoning to justify the leap from the given data of the problem to the above proportional relationship.* To understand what this proportional relationship means, let k be the common value of all these quotients. Then we have $f(n) = kn$ no matter what the integer n is. Thus proportional reasoning would include the recognition that this function $f(n)$, as a function defined on the positive integers, is a linear function. Since it is customary to regard this function as defined on the real numbers,[10] we have a linear function $f(x) = kx$ for all real numbers x. Furthermore, letting x be 1, then we have $f(1) = k$. *This is the statement that each person drinks k liters of water per day.*

It is important to emphasize this fact, namely, that a student well versed in proportional reasoning is assumed to be one who can infer from the given data that every person in the camping trip drinks the same amount of water per day, namely, k liters. In other words, we believe that students with conceptual understanding of proportional reasoning would be able to draw such a far-fetched conclusion. Indeed, even young kids can see that some people drink lots of water and others very little. This is not the kind of *mathematical reasoning* we want to postulate as desirable for students to acquire.

Let us consider the possibility that such faulty reasoning is solely the result of the faulty formulation of the problem. How would one formulate a problem having at least some contact with the "real world" along this line? A more responsible, and mathematically more accurate, formulation of the problem might read something like this:

> A group of 8 people are going camping for three days and need to carry their own water. They read in a guide book that 12.5 liters are needed for a party of 5 persons for 1 day. If one infers from the guide book that these figures provide a rough estimate of the amount of water consumed by a party of any size on any day, roughly how much water should they carry?

It goes without saying that the words, "provide a rough estimate", are nothing but code words for which students need precise and explicit explanations. One expects, therefore, that *students would receive instruction on reasoning of the following kind.* The key words

[10] In technical language, we extend in the obvious way the domain of f from the positive integers to the real numbers.

in the problem are those describing the amount of water consumed by a party of *"any size on any day"*. These are words that convey the *generality* of the message of the guide book (see the discussion of generality at the beginning of this article). If *any* 5 persons drink 12.5 liters per day, then students need to be made aware of the commonly accepted interpretation of this statement to mean that *any* person drinks roughly $\frac{12.5}{5} = 2.5$ liters on any day. Hence, for any positive integer n, n persons would drink, again roughly, $n \times 2.5 = 2.5n$ liters on a given day. Therefore it makes sense to define a function f so that $f(n)$ is roughly the amount of water n persons consume on a given day, and we saw that we had an expression for this f: $f(n) = 2.5n$ liters.

While the preceding discussion succeeds in making more sense of the conclusion that each person drinks 2.5 liters per day, the more important message is the need to make explicit, in one way or another, the underlying linear relationship in problems related to proportional reasoning. *One must set some ground rules so that students are not required to* ***guess*** *the linear relationship underlying the problem but that the linear relationship is made clear in some fashion.* We do not wish to enforce the rigid requirement that a linear function be handed to students in each problem of this type; such a requirement would be anti-educational. Rather, there should be universal recognition that linear relationships cannot be taken for granted, and students need explicit instructions as to when they take place.

Remember: mathematics has to be precise. Some may feel that when mathematics is taught with realistic contexts, students will build their own ideas and make sense of problems mathematically and make use of them to solve difficult problems. Perhaps, and perhaps not. But mathematics is not about saying things that would most likely be understood. Rather, it is about saying things in a way that would be *completely and unequivocally* understood. For this reason, if we want students to know that the "underlying relationship in a proportional situation" is a linear function, we must ensure that it is clearly understood as such. A correctly formulated mathematical problem is explicit about its assumptions, because being explicit about assumptions is a basic requirement of mathematics.

Once this point is understood, a classroom discussion of word problems related to proportional reasoning, whether correctly formulated or not, from the point of view of linear functions should be both revealing and rewarding.

From linear functions we go to quadratic ones. The graph of a linear function is a line, but what is the graph of a quadratic function? From our prior experience with quadratic equations, we first look at a special class of quadratic functions. The simplest is $f(x) = x^2$, and the starting point of our discussion is that the graph of $f(x) = x^2$ is known. The next is $f(x) = ax^2$ (for a real number a), then $f(x) = ax^2 + q$, and finally $f(x) = a(x+p)^2 + q$, for fixed constants a, p, and q. Let us concentrate on the last. From $f(-p+s) = f(-p-s)$ for any real number s, we see that the graph of f is symmetric with respect to the vertical line $x = -p$, and that the graph has its lowest (resp., highest) point at $(-p, f(-p))$, if $a > 0$ (resp., $a < 0$). Of course $f(-p) = q$. Moreover, the translation of the plane given by $(x, y) \longrightarrow (x+p, y-q)$ carries the graph of f onto the graph of the simple quadratic function $f_0(x) = ax^2$, so there is no doubt about the shape of the graph of f. So at least for simple quadratic functions expressible as $f(x) = a(x+p)^2 + q$, the graph is completely understood, and therewith, the function itself is completely understood. In fact, we can trivially read off the two points r_1 and r_2 on the x-axis at which the function f is equal to 0:

$$r_1,\ r_2 = -p \pm \sqrt{-\frac{q}{a}}$$

These are of course the roots of the quadratic equation $a(x+p)^2+q = 0$. (See the previous discussion on the relationship between the point where the graph of f crosses the x-axis and the solution of $f(x) = 0$.)

The fundamental theorem about quadratic functions is that, by the technique of completing the square, every quadratic function $f(x) = ax^2 + bx + c$ can be written in the form $f(x) = a(x + p)^2 + q$. In fact, simply let

$$p = \frac{b}{2a} \quad \text{and} \quad q = c - \frac{b^2}{4a}$$

It therefore follows from a comment in the last paragraph that the graph of the function $f(x) = ax^2 + bx + c$ can be made to coincide with the graph of $f_0(x) = ax^2$ by the use of a translation. Moreover, we also know from the last paragraph that the graph of $f(x) = ax^2+bx+c$ has an axis of reflection symmetry along the vertical line $x = -p = -\frac{b}{2a}$, and its lowest point (respectively, highest point) if a is positive (respectively, negative) is

$$(-\frac{b}{2a},\ c - \frac{b^2}{4a})$$

Thus by use of completing the square, one achieves a more conceptual way to look at the graph of f as regards shape and location.

Using the above expression of the zeros r_1 and r_2 in terms of p and q, we get immediately that $f(r_1) = f(r_2) = 0$ for

$$r_1,\ r_2 = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

Of course this is just another way of saying that r_1 and r_2 are the roots of the quadratic equation $ax^2 + bx + c = 0$.

The preceding discussion reveals that the technique of completing the square is much more than a skill to solve quadratic equations. Rather, it is the key to the understanding of quadratic functions in general.

Having quadratic functions available introduces still more word problems to the curriculum, especially those about falling objects in a gravitational field and certain work problems which are intractable up to this point. For example: Two workmen, painting at a constant rate, can paint a house together in 6 days. In how many days can each paint it alone if it takes one of them 3 days longer than the other to get it done?

The theory of quadratic polynomial functions, as presented above, is nearly complete, with one small piece missing: what happens when $b^2 - 4ac < 0$? In the next section, even this missing piece will be put in place. For polynomial functions in general, there is nothing as complete and simple. What one can hope to do is to sample a few polynomial functions of degree > 2, a few others which are not polynomial functions, and then make an effort to understand two new classes of functions: exponential and logarithmic functions. We will also make some passing comments on another class of functions that are important in the sciences: periodic functions. Let us first take up polynomial functions, but with one caveat that some new information will be withheld until we come to the fundamental theorem of algebra.

The simplest polynomial functions of higher degrees are those of the form $f_n(x) = x^n$, where n is a whole number > 1. The similarity in the shapes of the graphs of f_n to that of x^2 for all even n, and to that of x^3 for all odd n should be noted. Also noteworthy are the symmetry properties: for even n, the graph of f_n has a reflection symmetry with respect to the y-axis, and for odd n it has a radial symmetry with respect to the origin in the sense that a point (a, b) is on the graph if and only if $(-a, -b)$ is on the graph. In algebraic terminology, if n is even, f_n is an **even function** in the sense that $f_n(-x) = f_n(x)$, and if n is odd, f_n is an **odd function** in the sense that $f_n(-x) = -f_n(x)$. Through the plotting, point by point, of many examples, students can see that the behavior of the graph of an odd degree polynomial function on the positive x-axis differs from that on the negative x-axis: one gets higher and higher and the other, lower and lower. The explanation of this phenomenon in terms of a particular technique of factorization, viz.,

$$a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 = a_n x^n \left(1 + \frac{\frac{a_{n-1}}{a_n}}{x} + \cdots + \frac{\frac{a_1}{a_n}}{x^{n-1}} + \frac{\frac{a_0}{a_n}}{x^n} \right),$$

should be taught in place of hand-waving. With the explicit introduction of the intermediate theorem (without proof, of course), one extracts from this phenomenon the fact that, although even degree polynomial functions do not always have a zero, the odd degree ones always do.

The graph of the **absolute value function** $|x|$, and the graph of the **step function** $s(x)$ so that

$$s(x) = n \quad \text{for} \quad n \leq x < (n+1), \quad n \text{ an integer}$$

should also be singled out.

Another class of functions whose graphs are interesting are the **rational functions**, i.e., those which are quotients of polynomials, $\frac{f(x)}{g(x)}$, where $f(x)$ and $g(x)$ are polynomials. Observe that the domain of definition of a rational function is in general not the number line because of the zeros of the denominator. The simplest among these is the function $h(x) = \frac{1}{x}$. Its graph exhibits a generic phenomenon about rational functions: it has a horizontal **asymptote** (x-axis in this case) and a vertical asymptote (y-axis in this case). More complicated rational functions have slant asymptotes. For example, $f(x) = \frac{2x^3-1}{x^2+1}$ has the line $y = 2x$ for the asymptote. Needless to say, the concept of a **limit** should only be introduced informally in this context for the purpose of defining the asymptotes. Asymptotes are an integral part of the study of rational functions and they add a new element to the study of algebra.

Before one can tackle exponential functions, one has to discuss with care the concept of rational exponents. Students should be exposed to the underlying *mathematical* reason for the definitions of negative exponents and fractional exponents, which is to be able to extend the basic laws of exponents

$$a^m a^n = a^{m+n} \quad \text{and} \quad (a^m)^n = a^{mn}$$

from *positive integers* m and n to all rational numbers m and n. There is a need to emphasize the genesis of these definitions, because puzzlement on this subject appears to be widespread at the moment. Incidentally, before fractional exponents can be defined, i.e., the fact that $a^{m/n}$ is the m-th power of the n-th root of a, a careful discussion of the existence of the (positive) n-th root is necessary. As in the case of the square root, the relevant theorem is that, given a positive number a and a positive integer n, there is one and only one positive number b so that $b^n = a$. This b is called the (positive) ***n*-th root** of a, to be denoted by $\sqrt[n]{a}$. While the existence of $\sqrt[n]{a}$ must await a thorough study of the real numbers, students should learn how to prove the uniqueness statement. The most basic fact governing the operations with n-th roots is $\sqrt[n]{a}\,\sqrt[n]{b} = \sqrt[n]{ab}$, for all positive a and b, and this too should be proved because the proof introduces students to a typical bit of abstract thinking. Compare the earlier remark about the square root.

A word of caution is that the same laws for *rational* exponents, i.e.,

$$a^s a^t = a^{s+t} \quad \text{and} \quad (a^s)^t = a^{st}$$

where s and t are now arbitrary rational numbers, are excruciating to prove in general. A discussion of such laws in school algebra, therefore, would make more sense if it is focused on a few concrete examples with the intention of making these laws seem reasonable. Explain that in calculus, all exponents, rational or irrational, will be defined and the laws of exponents will be proved in one fell swoop rather than piecemeal.

With the availability of the concept of the n-th root of a positive number and the laws of exponents, one can now introduce **radical expressions** as numbers created when we apply to a collection of numbers x, y, etc. not only the existing arithmetic operations on numbers but also the new operation of taking the r-th power of a positive number, where r is any rational number. It is only when rational exponents are firmly understood that radical expressions can be taught as something more than rote procedures. Incidentally, students should be exposed to the graphs of $\sqrt{x}$ (defined only on $[0, \infty)$) and $\sqrt[3]{x}$ (defined on the x-axis).

Given a positive a, we can now define the **exponential function** $f_a(x) = a^x$. The number a is the **base**. A major reason for the introduction of rational exponents is precisely to make sense of this definition of an exponential function: we now know what $f_a(x)$ is for all *rational* values of x, and for the purpose of school mathematics, this knowledge is sufficient (recall FASM). One should emphasize the shape of the graph of f_a: it always passes through $(0, 1)$, but is above the x-axis and rises steeply to the right if $a > 1$, and slopes down to the x-axis to the right if $a < 1$. Contrast this with the graph of x^a. The number e can be introduced informally at this point in terms of the slope of the tangent to the graph at $(0, 1)$.

Before introducing logarithms, two things need be done. One is a careful explanation of the **composition** of functions, leading to the concept of **inverse functions** and the comparison of the graphs of a function and its inverse function. Both concepts are difficult for students, and the difficulty would likely be exacerbated by the lack of time. This may be a good reason to cut back on the number of topics in a second-year course in algebra.

The other is the historical reason which led to the discovery of logarithms. It should be mentioned in an algebra class not only because it is interesting history, but also because it gives an excellent motivation for the definition of the logarithm.

Let us confine ourselves for the moment only to numbers which can be expressed as an integer power of 3, i.e.,

$$\ldots, \frac{1}{9}(=3^{-2}),\ \frac{1}{3}(=3^{-1}),\ 0(=3^{0}),\ 3,\ 9(=3^{2}),\ 27(=3^{3}),\ 81(=3^{4}), \ldots$$

Notice that each such number is identified *uniquely* with its exponent as a power of 3, i.e., once 3 is fixed, then $\frac{1}{729}$ is uniquely identified with -6 (because $\frac{1}{729} = 3^{-6}$), 81 is uniquely identified with 4 (because $81 = 3^4$), 177147 is uniquely identified with 11 (because $177147 = 3^{11}$), etc. Let us devise a notation to indicate this identification: write $L(x)$ for the exponent of 3 when x is expressed as a power of 3. So

$$L(\frac{1}{729}) = -6, \quad L(81) = 4, \quad L(177147) = 11, \quad L(3^n) = n.$$

The law of exponents that $3^m\, 3^n = 3^{m+n}$ is now expressed in the new notation as

$$L(ab) = L(a) + L(b).$$

For example,

$$L(81 \times 177147) = L(81) + L(177147)$$

because

$$L(3^4 3^{11}) = L(3^{4+11}) = 4 + 11 = L(3^4) + L(3^{11})$$

Now suppose we want to find the product of two such numbers, say 177147×243. We can directly multiply, of course. But four centuries ago, John Napier (1550–1617) made the observation that, *by expressing a number in terms of the exponent of a fixed number* (which is 3 in our case), one could convert multiplication to addition. In greater detail, Napier's observation is the following. His "fixed number" was essentially e, but if he had used 3 instead of e, he would have had a table of the values of the powers of 3, including the following:

n	3^n	n	3^n	n	3^n
1	3	9	19683	17	129 140 163
2	9	10	59049	18	387 420 489
3	27	11	177147	19	1 162 261 467
4	81	12	531441	20	3 486 784 401
5	243	13	1594323	21	10 460 353 203
6	729	14	4782969	22	31 381 059 609
7	2187	15	14348907	23	94 143 178 827
8	6561	16	43046721	24	282 429 536 481

Then he would be able to use $L(177147)$ and $L(243)$ to do the multiplication of 177147×243 effortlessly: we have from the table that $177147 = 3^{11}$ and $243 = 3^5$, so that

$$L(177147 \times 143) = L(3^{11}\, 3^5) = L(3^{11+5}) = \mathbf{11} + \mathbf{5} = 16$$

But from the table, we see that $L(43046721) = 16$. Since as we observed above that 43046721 is the only number identified with 16, we get, for free,

$$177147 \times 143 = 43046721$$

provided we had the foresight to compile the table. We said we got the result "for free", because the only computation we did was to add $11 + 5$. *Addition is much simpler than multiplication.*

This discussion would seem to be too restrictive to be of interest. After all, the integer powers of 3 are a very small collection of numbers. What saves the day is the fact that *every positive real number* turns out to be expressible as a power of 3 with a real exponent. The same principle then shows that if we compile a table of the powers of 3 for a sufficiently many powers, then we would be able to multiply any two positive numbers, at least approximately, by use of such a table. In fact, there is no need to limit the discussion to using 3 as the base; everything that has been said so far holds for any base. Historically, base 10 was used, and the L in that case is called the **common logarithm**, denoted by $\log_{10}$. If 3 is used, as is the case at present, then this L would be written as $\log_3$. As mentioned above, Napier essentially used e as the base, and $\log_e$, or more simply log in advanced mathematics, is called the **natural logarithm**.[11] For over three centuries, tables of logarithms saved scientists countless hours in computations with data.

With the advent of computer software and calculators, one may legitimately ask what is the point of learning about the logarithm now? Many reasons, among them the following two. One is that multiplication is still more complicated than addition, so if a function log can convert multiplication into addition, in the sense of

$$\log ab = \log a + \log b \quad \text{for all positive } a,\ b$$

(i.e., $L(ab) = L(a) + L(b)$) it is automatically worth knowing. The other reason is that both exponential functions and logarithms figure prominently not just in mathematics but in all the sciences. There is no way one can avoid either.

[11] Engineers and some physicists denote the natural logarithm by ln.

Without any knowledge of the historical background of the logarithm, many students have come to regard it as another function they need to learn to pass an exam, with no appreciation whatsoever of the almost magical property that $\log x$ changes multiplication to addition.

Fixing a base a, $a > 0$ and $a \neq 1$. Then the functions $f(x) = \log_a x$ and $g(x) = a^x$ are inverse functions, the domain of f being all positive numbers. Thus

$$\log_a a^x = x \quad \text{for all } x$$

and

$$a^{\log_a t} = t \quad \text{for all positive } t$$

These two relations, which characterize the fact that $\log_a$ and a^x are inverse functions, are the key to solving equations of the following type: find x so that $11^{2x-1} = 5.8$. Equations of this type come up in applications, because nature seems to dictate that natural growth and decay processes be modeled by exponential functions, more specifically, by e^x and variants thereof. There is potential for real excitement here as simple carbon dating problems can be discussed in this context.

Problems about changing bases in logarithms are good for testing students' basic understanding of the definitions, but should not be elevated to the status of a major topic. (Compare the discussion about factoring quadratic polynomials.)

The last collection of functions to receive attention in school algebra are from outside algebra: the trigonometric functions. These are functions that are initially defined only for $0 \leq x \leq \frac{\pi}{2}$ (we use radian measure), so the first order of business is to extend the definition of sine, cosine to $0 \leq x \leq 2\pi$, and then to all values of x by demanding **periodicity of period 2π**, namely, $\sin(x \pm 2\pi) = \sin x$, and the same for cosine. Thus sine and cosine become defined on the whole number line. They turn out to be the prototypical periodic functions, because in advanced mathematics one shows that any function f which satisfies $f(x \pm 2\pi) = f(x)$ for all x can be *expressed in terms of sine and cosine* in a precise sense. The restriction to 2π as the period is more apparent than real because, if a positive number c is given, then the function $h(x) = a\sin(2\pi x/c)$ is periodic of period c (i.e., $h(x \pm c) = h(x)$ for all x), and has maximum value $|a|$ and minimum value $-|a|$. Students should have instant recall of the graphs of sine and cosine. In particular, sine is increasing in the interval $[-\frac{\pi}{2}, \frac{\pi}{2}]$ and therefore has an inverse function there. Similarly, cosine is decreasing on $[0, \pi]$ and it too admits an inverse function there.

One of the reasons that sine and cosine are important in mathematics and science is that nature is full of periodic phenomena.

Once sine and cosine are fully understood, the study of the other four trigonometric functions becomes fairly routine. Tangent is the most important of these four and its graph too should be accessible to instant recall. Note that tangent is not defined at all odd integer multiples of $\frac{\pi}{2}$.

We have mentioned the use of exponential and logarithmic functions to model growth and decay and the use of sine and cosine to model periodic phenomena. The use of linear functions to model certain observable phenomena should also be mentioned in this connection. Generically, some observational data, when properly graphed, seem to suggest a linear relationship although the data do not appear to be strictly linear due to inevitable observational inaccuracies. In such situations, one employs the method of **least squares** to arrive at a linear function that is a "best fit" for the data. The details of the method are beyond the level of school mathematics, but the main idea of the method can nevertheless be conveyed in a qualitative way. The use of computer software, or even a programmable calculator, to produce the "line of best fit" may be used, with guidance, to advantage to give an intuitive understanding of this process.

Algebra of Polynomials

Algebra at a more advanced level becomes an abstract study of structure. School algebra, at some point, should introduce students to such abstract considerations. The study of polynomials provides a good transition from algebra as generalized arithmetic to abstract algebra. Rather than considering a polynomial as the sum of multiples of powers of a *number* x, we now consider sums of multiples of powers of a *symbol* X,

$$f(X) = a_nX^n + a_{n-1}X^{n-1} + \cdots + a_1X + a_0$$

where the a_i's are constants $(i = 0, 1, \ldots, n)$, and $a_n \neq 0$. To avoid confusion with the polynomials we have been working with so far, we will call such an $f(X)$ a **polynomial form**, but will continue to refer to n as its **degree**, the a_n, …, a_0 as its **coefficients**, and a_n as its **leading coefficient**. The term a_0 will also be referred to as its **constant term**. As a matter of convention, we will omit the writing of a term if its coefficient is 0. Thus $2X^2 + 0X + 3$ is abbreviated to $2X^2 + 3$. We *define* two polynomial forms to be **equal** if their corresponding coefficients are pairwise equal. In particular, two equal polynomial forms must have the same degree, because they must have the same leading coefficient. Note that the concept of the *equality of polynomial forms* is a matter of definition and is not subject to psychological interpretations. This should help clarify the current concern about the meaning of the equal sign in the mathematics education research literature.

We now single out a special case of the equality of polynomial forms for further discussion because it is a delicate point that can lead to confusion. Given a polynomial form $f(X)$, what does it mean to say $f(X) \neq 0$? In this instance, the "0" can only mean the "0 polynomial form", i.e., "0" here means the polynomial form whose coefficients are all equal to 0. Therefore, as polynomial forms, the statement $f(X) \neq 0$ means that the two polynomial forms $f(X)$ and 0 are not equal. From the preceding definition of the equality of polynomial forms, we see that $f(X) \neq 0$ means there is at least one coefficient of $f(X)$ which is nonzero. Thus $X^2 - 1 \neq 0$. The reason this may be confusing is that, if one is not careful, one might think of $X^2 - 1 \neq 0$ as the statement that the *polynomial* $x^2 - 1$ is never equal to 0 for any number x. Obviously the latter is not true, e.g., $x^2 - 1$ is equal to 0 when $x = \pm 1$. In this instance, one has to exercise care in distinguishing between a polynomial and a polynomial form.

We do not assume that we have any prior knowledge of the symbol X. The addition or multiplication of polynomial forms therefore becomes a matter of definition, i.e., it is up to us to specify how to do these arithmetic operations among polynomial forms because such an $f(X)$ is no longer a number. We do so in the most obvious way possible, which is to define addition and multiplication among polynomial forms by treating X as if it were a number, so that at least formally, dealing with polynomial forms does not introduce any surprises. The idea is so simple that, in place of the most general definition possible (which would require the use of symbolic notations not appropriate for school mathematics), it suffices to indicate what is intended by a typical example in each case. Thus let

$$g(X) = a_2X^2 + a_1X + a_0, \quad \text{and} \quad h(X) = b_3X^3 + b_2X^2 + b_1X + b_0,$$

where the a_i's and b_j's are numbers, and $a_2 \neq 0$, $b_3 \neq 0$. Then by definition, their sum is

$$g(X) + h(X) = b_3X^3 + (a_2 + b_2)X^2 + (a_1 + b_1)X + (a_0 + b_0)$$

and their product is

$$\begin{aligned} g(X)h(X) &= (a_2b_3)X^5 + (a_1b_3 + a_2b_2)X^4 + (a_0b_3 + a_1b_2 + a_2b_1)X^3 \\ &\quad + (a_0b_2 + a_1b_1 + a_2b_0)X^2 + (a_0b_1 + a_1b_0)X + a_0b_0. \end{aligned}$$

In other words, the product of two polynomial forms is obtained by multiplying out all possible terms and then collecting like terms by their powers. It is immediately seen that, *because the additions and multiplications among the coefficients are associative, commutative, and distributive,* so are the addition and multiplication of polynomial forms.

Why polynomial forms instead of just polynomials? This question cannot be answered satisfactorily in the setting of school mathematics. The idea roughly is that, since in a polynomial $f(x)$, the variable x plays the primary role whereas the (constant) coefficients play a subordinate role, it is conceptually more clear to disengage x from the coefficients altogether. By singling out x as a symbol X in this manner, we open the way for X to assume other values distinct from those of the coefficients. For example, in linear algebra, one allows X to be a square matrix to obtain the so-called *characteristic polynomial* of the matrix. Moreover, it will be observed that the addition and multiplication of polynomial forms depend only on the fact that the coefficients obey the associative, commutative, and distributive laws. Anything we say about polynomial forms that depends only on addition and multiplication therefore becomes valid not just for real numbers as coefficients, but also for any number systems that satisfy these abstract laws, such as the complex numbers that will be taken up presently. This is an example of the power of abstraction and generality in algebra.

In a limited way, we can illustrate the advantage of the abstraction by considering the problem of division among polynomials. Imitating the fact that the division of polynomial functions leads to rational functions, we introduce a **rational form** as any expression of the type

$$\frac{f(X)}{g(X)}\,, \quad \text{where } f(X) \text{ and } g(X) \text{ are polynomial forms, and } g(X) \neq 0.$$

(Recall that $g(X) \neq 0$ merely means that at least one coefficient of $g(X)$ is not equal to 0.) We also agree to identify every polynomial form $f(X)$ with the rational form $\frac{f(X)}{1}$. For example, 0 is identified with $\frac{0}{1}$. We again emphasize that as it stands, $\frac{f(X)}{g(X)}$ is just a formal expression, and it is up to us to give it meaning. First, what does it mean that two such expressions are *equal*? Given two rational forms $\frac{f_1(X)}{g_1(X)}$ and $\frac{f_2(X)}{g_2(X)}$, we define

$$\boldsymbol{\frac{f_1(X)}{g_1(X)} = \frac{f_2(X)}{g_2(X)}}$$

to mean

$$f_1(X)g_2(X) = f_2(X)g_1(X)$$

Thus, every rational form $\frac{0}{g(X)}$ is equal to 0. In other words, we rely on the cross-multiplication algorithm in fractions as a guide to *define* the equality of rational forms.

As a consequence, "equivalent fractions" is automatically valid among rational forms, in the sense that for all polynomial forms $f(X)$, $g(X)$, $h(X)$, we have

$$\frac{f(X)}{g(X)} = \frac{h(X)f(X)}{h(X)g(X)}$$

We also define the addition and multiplication of rational forms by imitating fractions:

$$\begin{aligned} \frac{f_1(X)}{g_1(X)} + \frac{f_2(X)}{g_2(X)} &= \frac{f_1(X)g_2(X) + f_2(X)g_1(X)}{g_1(X)g_2(X)} \\ \frac{f_1(X)}{g_1(X)} \cdot \frac{f_2(X)}{g_2(X)} &= \frac{f_1(X)f_2(X)}{g_1(X)g_2(X)} \end{aligned}$$

One then shows that these operations are well-defined, and that a rational form $\frac{f(X)}{g(X)}$ has the desired property of the division of $f(X)$ by $g(X)$, in the sense that

$$g(X) \cdot \frac{f(X)}{g(X)} = f(X)$$

Note also that every nonzero rational form $\frac{f(X)}{g(X)}$ has a multiplicative inverse $\frac{g(X)}{f(X)}$. This completes the analogy that the set of rational forms is to the set of polynomial forms as the rational numbers are to the integers.

It may be wise to soft-pedal the well-definedness of the addition and multiplication of rational forms in a school classroom and use the time instead on more substantive things, such as the division algorithm and its consequences (see below).

If we were constrained to discuss polynomials rather than polynomial forms, then we would be faced with the awkward situation regarding the domain of a rational function $\frac{f(x)}{g(x)}$: it is not the number line but the points on the number line outside the zeros of $g(x)$, so that the domain of a sum $\frac{f(x)}{g(x)} + \frac{u(x)}{v(x)}$ is the set of the point on the number line outside the zeros of both $g(x)$ and $v(x)$, etc. But for polynomial forms, we have seen from the above discussion concerning a polynomial form being nonzero that there is no such awkwardness.

The analogy of polynomial forms with whole numbers leads us to consider the analog of division-with-remainder among whole numbers in the context of polynomial forms. This is the important **division algorithm** for polynomial forms: *given any polynomial forms $f(X)$ and $g(X)$, there are polynomial forms $Q(X)$ and $r(X)$ so that*

$$f(X) = Q(X)g(X) + r(X)$$

where $r(X)$ is either 0 or has a degree $<$ the degree of $g(X)$. The reasoning is essentially a repackaging of the familiar procedure of long division among polynomials. This is a basic fact about polynomial forms in the school algebra curriculum. We note explicitly that for the division algorithm to be valid, it is essential that beyond the associative, commutative, and distributive laws, the coefficients have the property that a nonzero number has a multiplicative inverse.

If c is a number and if we let $g(X) = (X - c)$ in the division algorithm, then we see immediately (because the degree of $g(X)$ is 1 so that $r(X)$, being of degree 0, must be a number) that $f(c) = 0$ if and only if $r(X) = 0$. Thus $f(c) = 0$ if and only if $f(X) = Q(X)(X - c)$, where degree of $Q(X)$ is 1 less than the degree of $f(X)$. We put this fact in the context of solving equations by defining a number c to be a **root of the polynomial form** $f(X)$ if $f(c) = 0$. Then repeated applications of the preceding fact yields: *if a polynomial form $f(X)$ has degree n, then it has at most n roots.* Along this line, the *rational root theorem* for equations with integer coefficients should be mentioned, although care must be exercised in the teaching of this theorem, in the following sense. Because the needed divisibility property of whole numbers for the proof of the rational root theorem is generally not proved in the earlier grades, this theorem tends to be offered for rote memorization without proof. As a result, a common pitfall is to mistake the theorem to be a test of the existence of *all real roots* rather than merely a test of the existence of *rational* roots.

Many polynomial equations do not have any roots among *real numbers*, e.g., $x^4+1 = 0$. By extending the real numbers to complex numbers, the situation changes drastically. It is easy to introduce complex numbers if students are used to the number line. What they have seen thus far is that it is possible to add and multiply any two points on the x-axis of the coordinate plane. Now they learn to do arithmetic with points *in the coordinate plane*, as follows: for all real numbers a, b, c, d, we define

$$\begin{aligned}(a, b) + (c, d) &= (a + c, b + d)\\ (a, b) \cdot (c, d) &= (ac - bd, ad + bc)\end{aligned}$$

One then proves that the points in the plane can be added, subtracted, multiplied, and divided in exactly the same way as real numbers. In other words, addition and multiplication of points in the plane obey the associative, commutative, and distributive laws, and for any nonzero complex number z, there is a complex number z^{-1} so that $zz^{-1} = z^{-1}z = 1$. The resulting number system is called the **complex numbers**, to be denoted by $\mathbb{C}$. A more common notation for complex numbers is to write $a+ib$ for (a, b), so that $i = (0, 1)$, by definition, and so that all real numbers t $(= t + i0)$ are just points $(t, 0)$ on the x-axis. The definition of multiplication among complex numbers of course implies that $i^2 = -1$, i.e., $(0,1)\cdot(0,1)=(-1,0)$. Any two complex numbers of the form $a+ib$ and $a-ib$ are said to be **conjugates** of each other. Notice that $(a+ib)(a-ib)=a^2+b^2$, which is *real*. *Notation*: $\overline{a+ib}=a-ib$.

The representation of complex numbers in terms of polar coordinates and de Moivre's theorem are essential ingredients even in a short account of complex numbers. The latter provides a straightforward method of taking **an *n*-th root** of a nonzero complex number z, i.e., to find a complex number w so that $w^n=z$. This discussion necessarily brings in the sine and cosine functions, thereby providing a nice tie-in with the earlier discussion of trigonometric functions. Note that de Moivre's theorem provides n such n-th roots, which is in contrast with the case of real numbers where we can specify a unique positive n-th root of any positive number.

It was mentioned that the concept of a polynomial form can be expanded to include any coefficients whose addition and multiplication obey the associative, commutative, and distributive laws. We do so now by allowing the coefficients of polynomial forms to be complex numbers, and call such polynomial forms **complex polynomial forms**. The major reason for the introduction of complex numbers can now be stated:

Fundamental Theorem of Algebra *Every complex polynomial form of positive degree has a complex root.*

The proof of this theorem is beyond the level of school mathematics, but students in algebra can achieve a firm grasp of the significance of this theorem by exploring its implications. The first consequence is to expand the previous argument using the division algorithm to conclude that every complex polynomial form of degree n can be expressed as a product

$$a(X-r_1)(X-r_2)\cdots(X-r_n)$$

where $a, r_1, \ldots, r_n$ are complex numbers. In particular, every complex polynomial form of degree n has *exactly* n roots (counting repeated roots). We note that this result depends on the validity of the division algorithm for complex polynomial forms, which in turn relies on the fact that any nonzero complex number has a multiplicative inverse. In the case of a complex polynomial form of degree 2, one can derive this result without invoking the Fundamental Theorem of Algebra. Indeed, the usual derivation by completing the square and the fact that *every* nonzero complex number has exactly two complex roots lead to an expression of the two roots by the quadratic formula. Incidentally, this shows that every real quadratic polynomial *always* has two roots, if complex numbers are allowed.

Suppose the coefficients of a polynomial form $f(X)$ are real numbers. It can be considered a complex polynomial form, of course, and therefore it is equal to a product $a(X-r_1)(X-r_2)\cdots(X-r_n)$ as before. But since the coefficients of $f(X)$ are real, a basic theorem of school algebra states that the roots r_1, ..., r_n must come in conjugate pairs. The proof of this fact, assuming the Fundamental Theorem of Algebra, is very instructive and should be mastered by every student. Because $(X-z)(X-\overline{z}) = X^2-(z+\overline{z})X+z\overline{z}$, and the coefficients of the latter are all real, we have proved that *every polynomial form with real coefficients is the product of* real *linear polynomial forms and* real *quadratic polynomial forms without real roots.*

We have thus far concentrated on polynomial forms with one symbol X. There is no reason not to consider polynomial forms in more than one symbol. A case in point is the very natural question of whether there is a formula for $(X+Y)^n$, where X and Y are two symbols and n is a positive integer. The answer to this question is given by the so-called *Binomial Theorem.*

The main impetus behind this question is the simple identity

$$(X+Y)^2 = X^2 + 2XY + Y^2.$$

A little bit more labor gives

$$(X+Y)^3 = X^3 + 3X^2Y + 3XY^2 + Y^3.$$

If one is persistent and computes the 4-th, 5-th, and even 6-th powers of $X+Y$, one would perceive a certain pattern and come up with a guess. A legitimate approach to the binomial theorem in school is to take for granted that such a guess has been made and proceed to prove it. Define the **binomial coefficients** for whole numbers $0 \le k \le n$ by

$$\left(\begin{array}{c} n \\ k \end{array} \right) = \frac{n!}{k!(n-k)!} ,$$

where by definition, $0! = 1$, and $n! = 1 \cdot 2 \cdot 3 \cdots (n-1) \cdot n$ for $n > 0$. Then:

Binomial Theorem *For all integers $n \ge 1$,*

$$(X+Y)^n = X^n + \left(\begin{array}{c} n \\ 1 \end{array} \right) X^{n-1}Y + \left(\begin{array}{c} n \\ 2 \end{array} \right) X^{n-2}Y^2 + \cdots + \left(\begin{array}{c} n \\ n-1 \end{array} \right) XY^{n-1} + Y^n$$

There are many ways to prove this theorem, and one way is to introduce the principle of **mathematical induction** and use it to give a proof. The teaching of mathematical induction should explain and stress the intuitive idea behind it rather than make it a mechanical procedure. It is also important to emphasize the fact that mathematical induction can be used only when the correct guess of a formula has been made. It cannot be used to deduce the formula. Another ingredient of the proof of the Binomial Theorem is the identity:

$$\left(\begin{array}{c} n \\ k-1 \end{array} \right) + \left(\begin{array}{c} n \\ k \end{array} \right) = \left(\begin{array}{c} n+1 \\ k \end{array} \right)$$

This may be proved directly as an exercise in the addition of fractions. Once this identity is available, one must mention the construction of Pascal's triangle on the basis of this identity. The proof of the Binomial Theorem then becomes a good exercise in mathematical induction.

As further illustration of the use of mathematical induction, one may use it to re-prove the formulas for the arithmetic series and geometric series.

Another proof of the Binomial Theorem can be given by considerations of permutations and combinations. We first give a different interpretation of the binomial coefficients, which can be proved by standard reasoning:

$$\left(\begin{array}{c} n \\ k \end{array} \right) = \text{ the number of } k\text{-element subsets of } \{1, 2, \ldots, n\}$$

Now think of $(X+Y)^n$ as the multiplication of n factors each equal to $(X+Y)$:

$$\underbrace{(X+Y)(X+Y)\cdot\cdots\cdot(X+Y)}_{n}$$

When the multiplication is carried out by the distributive law, each term will contain k Y's (and consequently $n-k$ X's), where k is equal to 0, 1, 2, 3, ..., $n-1$, n in succession. For a fixed k, we want to collect all the terms containing k Y's, and there are a total of $(\begin{array}{c} n \\ k \end{array})$ such terms because each such term comes from any k of the factors $(X+Y)$ and there are $(\begin{array}{c} n \\ k \end{array})$ ways to pick these k factors from the totality of n such factors. Thus when collecting these terms, we get

$$\left(\begin{array}{c} n \\ k \end{array} \right) X^{n-k} Y^k$$

This is then the Binomial Theorem.

Combinatorics and Finite Probability

The last proof of the preceding section naturally leads to *elementary* considerations of permutation and combination. The basic problems can be formulated abstractly as the following four:

(a) How many ways are there to place k distinctly colored balls in n distinctly numbered boxes, so that each box holds only one ball?

Answer: $\dfrac{n!}{(n-k)!}$.

(b) How many ways are there of placing k colored balls into n numbered boxes, if each box can hold as many balls as we wish?

Answer: n^k

(c) How many ways are there to place k balls of the same color in n numbered boxes, so that each box holds only one ball?

Answer: $\binom{n}{k}$

(d) How many ways are there of placing k colored balls into n numbered boxes, where (let us say) p of the balls are of one color, q of the balls are of a second color, and r ($r = k - p - q$) of the balls are of a third color, if each box holds only one ball?

Answer: $\dfrac{\binom{n}{k}}{p!q!r!}$

Combinations and permutations in turn lead naturally to (finite) probability, e.g., if I place 5 distinctly colored balls into 9 numbered boxes at random (each box holding only one ball), what is the probability that there are balls in the first two boxes? The basic concepts of probability should be discussed, if only lightly, such as the concept of a sample space, the fact that probabilities are numbers between 0 and 1, the fact that the sum of the probability of an event and the probability of the event not happening is 1, and the difference between dependent and independent events.

References

[Kodaira 1] Kunihiko Kodaira, Editor. *Mathematics 1: Japanese Grade 10.* American Math. Soc., 1996.

[Kodaira 2] Kunihiko Kodaira, Editor. *Mathematics 2: Japanese Grade 11.* American Math. Soc., 1997.

[NRC2001] *Adding It Up.* National Research Council, Washington D.C., 2001.

[Wu] How to prepare students for algebra, *American Educator*, Summer 2001, Vol. 25, No. 2, pp. 10-17.
http://www.aft.org/pubs-reports/american_educator/summer2001/index.html

Harvard University
University of California at Berkeley