Introduction

A mathematician, a musician, and a psychologist walked into a bar …

Several years ago, before I had any thoughts of writing a book on the history of symbols, I had a conversation with a few colleagues at the Cava Turacciolo, a little wine bar in the village of Bellagio on Lake Como. The psychologist declared that symbols had been around long before humans had a verbal language, and that they are at the roots of the most basic and primitive thoughts. The musician pointed out that modern musical notation is mostly attributed to one Benedictine monk Guido d'Arezzo, who lived at the turn of the first millennium, but that a more primitive form of symbol notation goes almost as far back as Phoenician writing. I, the mathematician, astonished my friends by revealing that, other than numerals, mathematical symbols—even algebraic equations—are relatively recent creations, and that almost all mathematical expressions were rhetorical before the end of the fifteenth century.

"What?!" the psychologist snapped. "What about multiplication? You mean to tell us that there was no symbol for 'times'?"

"Not before the sixteenth… maybe even seventeenth century."

"And equality? What about 'equals'? the musician asked.

"Not before… oh… the sixteenth century."

"But surely Euclid must have had a symbol for addition," said the psychologist. "What about the Pythagorean theorem, that thing about adding the squares of the sides of a right triangle?"

"Nope, . . . no symbol for 'plus' before the twelfth century!"

A contemplative silence followed as we sniffed and sipped expensive Barolo.

As it turned out, I was not correct. And far, far back in the eighteenth century BC, the Egyptians had their hieroglyphical indications of addition and subtraction in glyphs of men running toward or away from amounts to be respectively added or subtracted. And from time to time, writers of mathematical texts had ventured into symbolic expression. So there are instances when they experimented with graphic marks to represent words or even whole phrases. The *Bakhshâlî* manuscript of the second century BC records negative numbers indicated by a symbol that looks like our plus sign. In the third century, Diophantus of Alexandria used a Greek letter to designate the unknown and an arrow-like figure pointing upward to indicate subtraction. In the seventh century, the Indian mathematician Brahmagupta used a small black dot to introduce the new number we now call "zero." And symbols were timidly beginning to find their way into mathematics by the second half of the fifteenth century. Of course, for ages, there have been the symbols that we use to designate whole positive numbers.

That night at the *enoteca*, I didn't know that my estimate for the adoption of symbols was premature by several centuries. Sure, Diophantus in the third century had his few designations; however, before the twelfth century, symbols were not used for operational manipulation at the symbolic level—not, that is, for purely symbolic operations on equations. Perhaps I should have pushed the edge of astonishment to claim, correctly, that *most* mathematical expressions were rhetorical before the sixteenth century.

Ever since that conversation, I have found that most people are amazed to learn that mathematics notation did not become really symbolic before the sixteenth century. We must also wonder: What was gained by algebra taking on a symbolic form? What was lost?

Traced to their roots, symbols are a means of perceiving, recognizing, and creating meaning out of patterns and configurations drawn from material appearance or communication.

The word "symbol" comes from the Greek word for "token," or "token of identity," which is a combination of two word-roots, *sum* ("together") and the verb *ballo* ("to throw"). A more relaxed interpretation would be "to put together." Its etymology comes from an ancient way of proving one's identity or one's relationship to another. A stick or bone would be broken in two, and each person in the relationship would be given one piece. To verify the relationship, the pieces would have to fit together perfectly.

On a deeper level, the word "symbol" suggests that, when the familiar is thrown together with the unfamiliar, something new is created. Or, to put it another way, when an unconscious idea fits a conscious one, a new meaning emerges. The symbol is exactly that: meaning derived from connections of conscious and unconscious thoughts.

Can mathematical symbols do that? Are they meant to do that? Perhaps there should be a distinction between symbols and notation. Notations come from shorthand, abbreviations of terms. If symbols are notations that provide us with subconscious thoughts, consider "+." Alone, it is a notation, born simply from the shorthand for the Latin word *et*. Yes, it comes from the "t" in *et*. We find it in 1489 when Johannes Widmann wrote *Behende und hubsche Rechenung auff allen Kauffmanschafft* (*Nimble and neat calculation in all trades*). It was meant to denote a mathematical operation as well as the word "and."

Used in an arithmetic statement such as $2 + 3 = 5$, the "+" merely tells us that 2 and 3 more make 5. But in the context of an algebraic statement such as $x^2 + 2xy + y^2$ it generally means more than just "x^2 and $2xy$ and y^2." The mathematician sees the +'s as the glue to form the perfect square $(x + y)^2$. Now surely the same mathematician would just as well see the "and" as the glue. Perhaps it may take a few more seconds to recognize the perfect square, but familiar symbols habitually provide useful associations when we are looking at one object while knowing that it has another useful form.

A purist approach would be to distinguish symbolic representation from simple notation. I have a more generous slant; numerals and all nonliteral operational notation are different, but still considered symbols, for they represent things that they do not resemble.

Read the statement 2 + 3 = 5 again. It is a complete sentence in mathematics, with nouns, a conjunction, and a verb. It took you about a second to read it and continue on. Unaware of your fact-checking processes, you believe it for many reasons, starting from what you were told as a young child and ending with a mountain of corroborating evidence from years of experience. You didn't have to consciously search through your mental library of truthful facts to *know* that it is true.

Yet there is a distinct difference between the writer's art and the mathematician's. Whereas the writer is at liberty to use symbols in ways that contradict experience in order to jolt emotions or to create states of mind with deep-rooted meanings from a personal life's journey, the mathematician cannot compose contradictions, aside from the standard argument that establishes a proof by contradiction. Mathematical symbols have a definite initial purpose: to tidily package complex information in order to facilitate understanding.

Writers have more freedom than mathematicians. Literary symbols may be under the shackles of myth and culture, but they are used in many ways. Emily Dickinson never uses the word "snake" in her poem "A Narrow Fellow in the Grass," thereby avoiding direct connections with evil, sneakiness, and danger, though hinting all the same. Joseph Conrad invokes all the connotations of slithering, sneaky evil in *Heart of Darkness* when describing the Congo River as "an immense snake uncoiled, with its head in the sea." It is also possible that a writer may use the word "snake" innocently, in no way meaning it as something unsuspected, crafty, or dangerous. It could be simply a descriptive expression, as in "the river wound around its banks like a snake." The writer may intend to invoke an image in isolation from its cultural baggage. This is tough—perhaps impossible—to do with words or expressions that are so often used figuratively.

Mathematicians use a lemma (a minor theorem used as a stepping stone to prove a major theorem) called the "snake lemma," which involves a figure called the "snake diagram"—it doesn't mean that there is anything sinister, crafty, or dangerous within, but rather that the figure just happens to look like a snake, again just a graphic description.

Human-made symbols of mathematics are distinct from the culturally flexible, emotional symbols found in music or from the metaphorical symbols found in poems. However, some also tend to evoke subliminal, sharply focused perceptions and connections. They might also transfer metaphorical thoughts capable of conveying meaning through similarity, analogy, and resemblance, and hence are as capable of such transferences as words on a page.

In reading an algebraic expression, the experienced mathematical mind leaps through an immense number of connections in relatively short neurotransmitter lag times.

Take the example of π, the symbol that every schoolchild has heard of. As a symbol, it is a sensory expression of thought that awakens intimations through associations. By definition, it means a specific ratio, the circumference of a circle divided by its diameter. As a number, it is approximately equal to 3.14159. It masquerades in many forms. For example, it appears as the infinite series

$$\pi = \frac{4}{1} - \frac{4}{3} + \frac{4}{5} - \frac{4}{7} + \frac{4}{9} \cdots,$$

or the infinite product

$$\pi = 2 \cdot \frac{2}{1} \cdot \frac{2}{3} \cdot \frac{4}{3} \cdot \frac{4}{5} \cdot \frac{6}{5} \cdot \frac{6}{7} \cdot \frac{8}{7} \cdot \frac{8}{9} \cdots,$$

or the infinite fraction

$$\pi = \cfrac{4}{1 + \cfrac{1^2}{3 + \cfrac{2^2}{5 + \cfrac{3^2}{7 + \ldots}}}}.$$

It frequently appears in both analytical and number theoretic computations. When she sees π in an equation, the savvy reader automatically knows that something circular is lurking behind. So the symbol (a relatively modern one, of course) does not fool the mathematician who is familiar with its many disguises that unintentionally drag along in the mind to play into imagination long after the symbol was read.

Here is another disguise of π: Consider a river flowing in uniformly erodible sand under the influence of a gentle slope. Theory predicts that over time the river's actual length divided by the straight-line distance between its beginning and end will tend toward π. If you guessed that the circle might be a cause, you would be right.

The physicist Eugene Wigner gives an apt story in his celebrated essay, "The Unreasonable Effectiveness of Mathematics in the Natural Sciences."[1] A statistician tries to explain the meaning of the symbols in a reprint about population trends that used the Gaussian distribution. "And what is this symbol here?" the friend asked.

"Oh," said the statistician. "This is pi."

"What is that?"

"The ratio of the circumference of the circle to its diameter."

"Well, now, surely the population has nothing to do with the circumference of the circle."

Wigner's point in telling this story is to show us that mathematical concepts turn up in surprisingly unexpected circumstances such as river lengths and population trends. Of course, he was more concerned with understanding the reasons for the unexpected connections between mathematics and the physical world, but his story also points to the question of why such concepts turn up in unexpected ways within pure mathematics itself.[2]

The symbol π had no meaning in Euclid's *Elements* (other than its being the sixteenth letter of the ancient Greek alphabet), even though the *Elements* contained the proof of the hard-to-prove fact that the areas of any two circles are to one another as the squares on their diameter.[3] The exceptionality of Greek mathematical thinking is in conceiving that there are universal truths that could be proven: that any circle is bisected by any of its diameters, that the sum of angles in any triangle is always the same constant number, that only five regular solids can exist in three dimensions. In book II, proposition 4, Euclid showed us how to prove what we might today think of as simple algebraic identities, such as $(a + b)^2 = a^2 + b^2 + 2ab$, but you will not find any algebraic symbols indicating powers (those little raised numbers that tell how many times to multiply a number by itself) or addition in his proposition or proof because his statements and proofs were, on the one hand, geometrical and, on the other, entirely in narrative form.

Diophantus of Alexandria was born more than five hundred years after Euclid. His great work, *Arithmetica*, gave us something closer to algebraic solutions of special linear equations in two unknowns, such as $x + y = 100$, $x - y = 40$. He did this not by using the full power of symbols, but rather by syncopated notation—that is, by the relatively common practice of the time: omitting letters from the middle of words. So his work never fully escaped from verbal exposition.[4] It was the first step away from expressing mathematics in ordinary language.

It is possible to do all of mathematics without symbols. In general, articles of law contain no symbols other than legalese such as "appurtenances," "aforesaid," "behoove"—words that few people would dream of using in anything other than a legal document. By tradition, and surely by design, law has not taken the symbolic road to precision. Words in a natural language such as English or Latin can present tight meaning, but almost never ironclad precision the way symbolic algebra can. Instead, written law relies heavily on intent, and expects loopholes to be found by those clever people who use them.

Imagine what mathematics would be like if it were still entirely rhetorical, without its abundance of cleverly designed symbols. Take a passage in al-Khwārizmī's *Algebra* (ca. 820 AD) where even the numbers in the text are expressed as words:

> If a person puts such a question to you as: "I have divided ten into two parts, and multiplying one of these by the other the result was twenty-one;" then you know that one of the parts is thing, and the other is ten minus thing.[5]

We would write this simply as: $x(10 - x) = 21$.

The language of the solution, as al-Khwārizmī wrote it, was specific to the question. There may have been a routine process, some algorithm, lurking behind the phrasing, but it would have taken work to bring it out, since al-Khwārizmī's *Algebra* is not particularly representative of the mathematics of his period.

Privately things may have been different. Thinking and scratch work probably would have gone through drafts, just as they do today. I have no way of knowing for sure, but I suspect that the solution was first probed on some sort of a dust board using some sort of personal notation, and afterward composed rhetorically for text presentation.

The sixth-century prolific Indian mathematician-astronomer Aryabhatta used letters to represent unknowns. And the seventh-century Indian mathematician-astronomer Brahmagupta—who, incidentally, was the first writer to use zero as a number—used abbreviations for squares and square roots and for each of several unknowns occurring in special problems. Both Aryabhatta and Brahmagupta wrote in verse, and so whatever symbolism they used had to fit the meter. On seeing a dot, the reader would have to read the word for dot. This put limitations of the use of symbols.[6] A negative number was distinguished by a dot, and fractions were written just as we do, only without the bar between numerator and denominator.

Even as late as the early sixteenth century, mathematics writing in Europe was still essentially rhetorical, although for some countries certain frequently used words had been abbreviated for centuries. The abbreviations became abbreviated, and by the next century, through the writings of François Viète, Robert Recorde, Simon Stevin, and eventually Descartes, those abbreviations became so compacted that all the once-apparent connections to their origins became lost forever.

In mathematics, the symbolic form of a rhetorical statement is more than just convenient shorthand. First, it is not specific to any particular language; almost all languages of the world use the same notation, though possibly in different scriptory forms. Second, and perhaps most importantly, it helps the mind to transcend the ambiguities and misinterpretations dragged along by written words in natural language. It permits the mind to lift particular statements to their general form. For example, the rhetorical expression *subtract twice an unknown from the square of the unknown and add one* may be written as $x^2 - 2x + 1$. The symbolic expression might suggest a more collective notion of the expression, as we are perhaps mentally drawn from the individuality of $x^2 - 2x + 1$ to the general quadratic form $ax^2 + bx + c$. We conceive of $x^2 - 2x + 1$ merely as a representative of a species.

By Descartes's time at the turn of the seventeenth century, rhetorical statements such as

> The square of the sum of an unknown quantity and a number equals the sum of the squares of the unknown and the number, augmented by twice the product of the unknown and the number.

were written almost completely in modern symbolic form, with the symbol ∞ standing for equality:

$$(x + a)^2 \infty x^2 + a^2 + 2ax$$

The symbol had finally arrived to liberate algebra from the informality of the word.

As with almost all advances, something was lost. We convey modern mathematics mostly through symbolic packages, briefcases (sometimes suitcases) of information marked by symbols. And often those briefcases are like Russian *matrioshka* dolls, collections of nested briefcases, each depending on the symbols of the next smaller *matrioshka*.

There is that old joke about joke tellers: A guy walks into a bar and hears some old-timers sitting around telling jokes. One of them calls out, "Fifty-seven!" and the others roar with laughter. Another yells, "Eighty-two!" and again, they all laugh.

So the guy asks the bartender, "What's going on?"

The bartender answers, "Oh, they've been hanging around here together telling jokes for so long that they catalogued all their jokes by number. All they need to do to tell a joke is to call out the number. It saves time."

The new fellow says, "That's clever! I'll try that."

So the guy turns to the old-timers and yells, "Twenty-two!"

Everybody just looks at him. Nobody laughs.

Embarrassed, he sits down, and asks the bartender, "Why didn't anyone laugh?" The bartender says, "Well, son, you just didn't tell it right…"

Mathematicians often communicate in sequentially symbolic messages, a code, unintelligible to the uninitiated who have no keys to unlock those briefcases full of meaning. They lose the public in a mire of marks, signs, and symbols that are harder to learn than any natural language humans have ever created.

More often, in speaking, for the sake of comprehension, they relax their airtight arguments at the expense of mildly slackening absolute proof. They rely on what one may call a "generosity of verbal semantics," an understanding of each other through a shared essence of professional expertise and experience independent of culture.

However, even with a generosity of verbal semantics, something beyond absolute proof is lost. Mathematics, even applied mathematics, physics, and chemistry, can be done without reference to any physically imaginable object other than a graphic symbol. So the difference between the physicist's rhetorical exposition and the mathematician's is one of conceptualization.

That might be why physicists have an easier time communicating with the general public; they are able to give us accounts of "stuff" in this world. Their stuff may be galaxies, billiard balls, atoms, elementary particles of matter and strings, but even those imperceptible strings that are smaller than 10^{-35} of a meter in ten-dimensional space are imagined as stuff. Even electric and magnetic fields can be imagined as stuff. When the physicist writes a book for a general audience, she starts with the advantage of knowing that every one of her readers will have experience with some of the objects in her language, for even her most infinitesimal objects are imaginable "things."

A mathematician's elements are somewhat more intangible. The symbol that represents a specific number N is more than just a notational convenience to refer to an indeterminate number. These days it represents an object in the mind with little cognate reference to the world—in other words, the N is a "being" in the mind without a definite "being" in the world. So the nonphysical object has an ontology through cognition. The mind processes a modern understanding of a number—say, three—just as it does for any abstraction, by climbing a few rungs of increasing generality, starting at a definite number of things within the human experience: three sheep in the field, three sheep, three living things, three things, . . . all the way up to "three-ness." Imagining physical objects decreases as generality increases. The mathematical symbol, therefore, is a visual anchor that helps the mind through the process of grasping the general through the particular.

This book traces the origins and evolution of established symbols in mathematics, starting with the counting numbers and ending with the primary operators of modern mathematics. It is chiefly a history of mathematical symbols; however, it is also an exploration of how symbols affect mathematical thought, and of how they invoke a wide range of enduring subconscious inspirations.

It is arranged in three parts to separate the development of numerals from the development of algebra. This was a difficult authoring decision based on fitting an acceptable symbol definition into the broader scope of notation, which includes that of both numerals and algebra. Each part has its separate chronology. Parts 1 and 2 are quasi-independent, but the reader should be aware that at early stages of development, both numerals and algebraic symbols progressed along chronologically entangled lines.