Signed numbers and signed letters in algebra

AMIR ASGHARI

The Glossary of the Common Core State Standards for Mathematics (2010) defines an integer by *a way that it can be expressed.*

> **Integer.** A number expressible in the form a or $-a$ for some whole number a.

This definition hints at three mathematical objects, a signed number, a signed letter and an integer, that it is useful to distinguish. A *signed number* is a specific natural number prefixed by the sign – or + [1]. A *signed letter* is a letter prefixed by the sign – or + to denote an unspecified negative or positive quantity. A letter may stand for a parameter or a variable, and I will use the expressions 'signed parameter' and 'signed variable' to reflect the role of the letter in the situation under discussion. In both signed numbers and signed letters, the + sign is often omitted (as it is in the Common Core definition), and here my focus is on the – sign. Signed numbers and signed letters have two parts, the number or letter that indicates a specified or unspecified number greater than 0, and the sign. An integer, on the other hand, is a unity, a number that, when added to another integer, gives the sum 0.

Thus for example, '−3' as a signed number is the number 3 prefixed by the sign – . But, as an integer, '−3' is the solution of ? + 3 = 0; "all we need to know about −3 is that when you add 3 to it you get 0" (Gowers, p. 26). Similarly, '$-a$' as a signed letter is the natural number a prefixed by the sign –. But, as an integer, '$-a$' is the solution of $? + a = 0$, and a itself might be less than 0.

The purpose of this short communication is to provide some historical evidence to show the role of signed letters in the development of algebra. In particular, some ingenious and correct ways that mathematicians used signed letters for expressing generality are shown, that now seem cumbersome and inefficient. It is suggested that we should be more aware and explicit about the use of signed letters and numbers in schools today.

Negative numbers at the advent of algebra

The historical rejection of integers as numbers is well known. Eventually, their utility was one of the most important reasons for their acceptance and development. For example, as Hefendehl-Hebeker (1991) writes "using negative numbers one can quickly and effectively solve problems that make sense only for positive numbers and have only positive solutions" (p. 32). An important factor in the gradual acceptance of negative numbers was due to the usefulness of the rules of signs applied to signed numbers and signed letters. According to Nathaniel Hammond's 1742 *The Elements of Algebra in a New and Easy Method* [2] the rules are as follows :

> When the quantities to be multiplied have like signs, that is, they are both affirmative, or both negative, then set or join the letters together, and to them prefix the sign +, which will be the product required. (Art. 9, p. 18)
>
> When the signs of the two quantities that are to be multiplied are one affirmative and the other negative, then multiply the quantities as before directed, but to their product prefix the negative sign, or – . (Art. 16, p. 26)

With a slight change of language (*e.g.,* using 'positive' instead of 'affirmative'), this is "the rule for multiplying" given by the most popular dedicated educational site in the UK [3]:

> When the signs are different the answer is negative.
>
> When the signs are the same the answer is positive.

Signed numbers and signed letters with the correct use of the rules of signs suffice for getting the correct answer in calculations and working with algebraic expressions. They even work well for finding the solutions of equations, but they are inefficient for expressing the equations generally for which we need to use parameters.

Signed parameters

> We have got this equation $x^2 + 3x - 10 = 0$, whence $x^2 + 3x = 10$ (Hammond, Art. 70, p. 249).

For us, $x^2 + 3x - 10 = 0$ is a special case of the general form of quadratic equations $x^2 + bx + c = 0$, with $b = 3$ and $c = -10$. But for Hammond, it is a special case of $x^2 + bx = c$ with $b = 3$ and $c = 10$. The other two forms [4] of quadratic equations are $x^2 - bx = c$ and $x^2 - bx = -c$.

With *signed parameters* (*i.e.,* signed letters to represent the coefficients), not only the general expressions of the quadratic equations but also the relevant theorems are divided into signed cases. For example, instead of the general theorem that the sum of the roots of $x^2 + bx + c = 0$ is $-b$, there are two theorems :

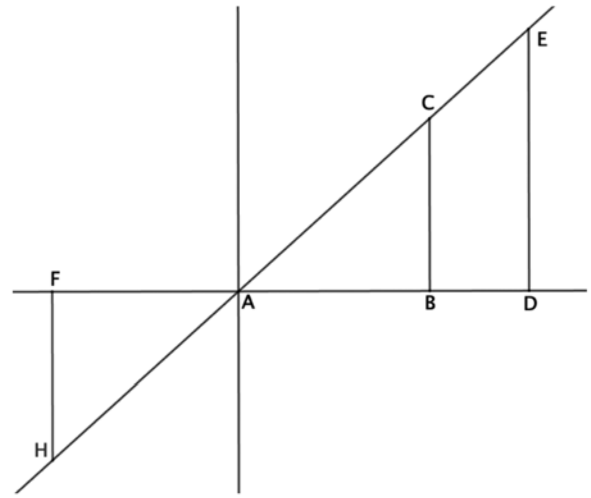
> If the co-efficient of [x] has the sign, + the sum of both the roots will be the same as the co-efficient, but will have the sign –.
>
> And, if the co-efficient of [x] has the sign –, [...] then the sum of both the roots or values, will be the same as the co-efficient, but will have the sign +.
>
> Therefore having found any one root, the other is easily found. (Art. 62, p. 188)

Let us consider the form where the co-efficient of x has the sign +. To find the other value of the unknown quantity, the first value is added to the co-efficient of x, "and to their sum prefix the sign –"(Art. 63, p. 188). For example, having found $x = 5$ as one of the roots of $x^2 + 10x = 75$, we add 5 to 10. The sum is 15. "Prefix to this 15 the sign – and this is the other value of x, that is $x = -15$." (Art. 63, pp. 188–189).

It would be constructive to compare Hammond's approaches with ours. For us, the sum of the roots of $x^2 + bx + c = 0$, is $-b$. Let the roots be x_1 and x_2, we can write $x_1 + x_2 = -b$. Now it does not matter if the root we have is positive or negative. Since x_1 admits both positive and negative numbers, we substitute the found root into $x_1 + x_2 = -b$ for x_1 and knowing b (that also could be positive or negative), we find x_2. This line of thought, that may seem very natural to the reader, is beyond Hammond's *Elements of Algebra.* Hammond's letters cannot admit varying negative quantities. They can represent *fixed unknown quantities* in equations. But, if –15, for example, is a solution to an equation, it must be "an *imaginary* value of *x,* it being absurd for a positive quantity to be equal to a negative one." (Art. 63, p. 189). Though it is "absurd" for the letter representing the root to be equal to a negative quantity, it is in line with the other rules of algebra, and hence, its acceptance is somehow justified: "we shall find this *imaginary* value of *x,* if we proceed by division according to the directions at Art. 61." (Art. 63, p. 189).

The whole process of finding the roots is based on the established rules of signs applying to signed numbers and signed letters. Though "absurd", the fixed value may turn out to be negative, but it is never *chosen* to be negative in the sense that a variable may admit negative numbers as well as positive. So for example, "No quantities connected by the sign + only [...], can be equal to nothing. That is, it cannot be [...] $a + b = 0$ (Hammond, Art. 77, p. 326).

Before reading the next section, please imagine yourself in the Hammond's time and try to draw the graph of the equation $y = -x$, that is $x + y = 0$.



Through the points A, C, draw an indefinite right line HE; this will be the locus of the two equations $y = \frac{ax}{b}$ and $y = -\frac{ax}{b}$. For, taking any line $AD = x$, and drawing DE parallel to BC, it will be $DE = \frac{ax}{b} = y$. And taking $AF = -x$, and drawing FH parallel to BC, it will be $FH = -\frac{ax}{b} = y$.

Figure 1. Agnesi's description of the graph of y = (a/b)x (redrawn and simplified, from Book I, Section III, Art 117; Colson, p. 93).

Signed variables

To observe a true use of letters as variables we have to move a few years forward from Hammond to 1748, the publication year of Maria Gaetana Agnesi's *Instituzioni Analitiche* (translated as *Analytical Institutions for the Use of Italian Youth* by John Colson, 1801 [5]). It includes some of the earliest examples of using co-varying quantities for expressing algebraic equations of curves, and "is significant for its clarity of exposition and its widespread influence as a textbook" (Boyer, 2012, p. 177). In it we can also find an explicit introduction to signed numbers and letters in algebra:

> Positive and negative quantities in algebra are distinguished by means of certain marks, or signs, which are prefixed to them. To positive quantities the sign +, or *plus,* is prefixed; to negative quantities the sign –, or *minus.* (Agnesi, 1748, Book I, Section I, Art. 3; p. 2 in Colson)

With $+a$ only denoting a positive quantity and $-a$ only a negative quantity, the general equation of a straight line (that is, for us $y = mx + b$), should be divided into different types, that according to Agnesi are as follows [6]: $y = mx$; $y = -mx$; $y = mx + c$; $y = -mx - c$; $y = mx - c$; $y = -mx + c$ (Book I, Section III, Art. 117; pp. 92-93 in Colson).

This is another example of signed parameters used for expressing a generality in *the absence of a letter admitting both positive and negative numbers.* But then in the equation of each of the lines above there are also two letters x and y that are not for representing fixed numbers. They are *variable quantities,* "because the value of one of the unknown quantities may be varied an infinite number of ways, so, in like manner, the value of the other may be as often varied" (Book I, Section III, Art. 111; p. 90 in Colson). The locus of each of the equation above is a right 'line', but not the one we are used to. The equation $y = mx$ is a ray in the first quadrant and $y = -mx$ is a ray in the third quadrant. In other words, to algebraically describe the line HE in Figure 1, we need two equations rather than one.

This is not, as Boyer suggests, "some of the old errors with respect to negative coordinates" (Boyer, 2012, p. 178). Rather, it is a *correct* way of dealing with the situation when the variable quantities are *signed variables.* Nowadays, it would be an error if our students assumed that $y = mx$ only represents positive y's and $y = -mx$ only negative y's. We expect them to see that the point (x, y) may lie in any of the quadrants, both x and y are unsigned variables admitting both positive and negative numbers, m is a parameter representing both positive and negative numbers, and hence, $y = mx$ is the general equation of lines passing through the origin.

School mathematics

We now present graphs of equations on the Cartesian plane with positive and negative numbers on the axes, and expect that our students will not experience difficulties accepting unsigned variables. After all, even a line as simple as $y = x$, assumes both positive and negative numbers for the variables involved. However, Christou and Vosniadou's (2005) study shows that students have a "strong tendency to interpret literal symbols as standing only for natural numbers" (p. 456). Even those few students who assigned a negative value to a literal symbol, say a, did it by "by putting a minus sign in front of it" (p. 455), changing it to $-a$.

Historically, the acceptance of unsigned variables was not easy. And it seems that even now it is not easy for our students. This difficulty may become less surprising if we consider that it is quite possible to do arithmetic, solve equations, and work with algebraic expressions, solely using signed numbers and signed letters (for this purpose, Hammond's book is still usable). And many students only encounter positive variables (mainly variables that admit natural numbers, *e.g.,* in so-called matchstick patterns) for a long time, before experiencing variables in general.

We should not assume that the transition to unsigned variables happens spontaneously for our students. The fact that variables can take negative values should be addressed directly. It may help students to see how unsigned variables simplify things, as they can use a *single* equation $y = mx + b$, for a line in the plane, while Agnes needed six.

Notes

[1] Signed numbers can also represent negative and positive rational or irrational numbers, but for simplicity I will at first consider signed numbers corresponding to integers.

[2] Available online at https://archive.org/details/elementsalgebra02hamm-goog. I have updated Hammond's notation slightly (*e.g.,* using x as the variable, and writing x^2 instead of xx) and changed the capitalisation of words to follow current standards.

[3] BBC Bitesize at https://www.bbc.co.uk/bitesize/guides/z77xsbk/revision/3

[4] There should be the fourth form $x^2 + bx = -c$. But Hammond does not discuss it. In cases where one of the roots is negative he justifies excluding it, so I believe that he was aware of the fourth case but chose to overlook it as it has only negative "absurd" roots.

[5] Available online at https://archive.org/details/analyticalinsti00masegoog

[6] Agnesi represented m by the fraction a/b and hence used c for the y-intercept.

References

Boyer, C.B. (2012) *History of Analytic Geometry.* Mineola, NY: Dover.

Common Core State Standards Initiative. (2010) *Common Core State Standards for mathematics.* Online at http://www.corestandards.org/Math/Content/mathematics-glossary/glossary/

Christou, K.P. & Vosnaidou, S. (2005) How students interpret literal symbols in algebra. In Bara, B.G., Barsalou, L. & Bucciarelli, M. (Eds.) *Proceedings of the 27th Annual Conference of the Cognitive Science Society,* pp. 453–458. Mahwah, NJ: Lawrence Erlbaum Associates.

Gowers, T. (2002) *Mathematics: A very short Introduction,* Vol. 66. Oxford: Oxford Paperbacks.

Hefendehl-Hebeker, L. (1991) Negative numbers: obstacles in their evolution from intuitive to intellectual constructs. *For the Learning of Mathematics* **11**(1), 26–32.