

The Equal Sign, Equations, and Functions, Draft 3

by Solomon Friedberg, February 23, 2009

An *equation* is a statement that two expressions are equal. A *function* is a rule that assigns to each eligible input a specific, single, output. Both can be described using the equal sign. Since we use equations to describe functions, it's easy to get confused about these concepts. Here we clarify this.

1. The Equal Sign and Equations

Students first meet the equal sign in problems for which “=” calls for a numerical evaluation. Thus a problem in the first grade could read

$$5 + 3 = _ _ _ .$$

Filling in the correct answer, $5 + 3 = 8$, gives an equation; the equal sign indicates that the numerical quantities on the two sides are equal.¹ An equation is *true* if the evaluation of the two sides gives the same value; otherwise it is false. An equation can be true even if both sides involve numerical operations. For example,

$$999 + 36 = 1000 + 35$$

is true, since both sides are the same quantity (1035); in addition, it suggests a way of adding 999 and 36 that is more efficient than the standard regrouping method. An equal sign may be used repeatedly to indicate that many quantities are all equal—but *only* when all the quantities are equal to each other. An example is

$$26 - 9 = (20 + 6) - 9 = (10 + 16) - 9 = 10 + (16 - 9) = 10 + 7 = 17.$$

It is often preferable to write such an equation in the form

$$\begin{aligned} 26 - 9 &= (20 + 6) - 9 \\ &= (10 + 16) - 9 \\ &= 10 + (16 - 9) \\ &= 10 + 7 \\ &= 17. \end{aligned}$$

Students are sometimes reluctant to use the equal sign in their work—if they are wrong, they are more likely to have points deducted, so it is better to be vague. However, the

¹A more careful form of the problem “ $5 + 3 = _ _ _$ ” would be “Evaluate the sum $5 + 3$ and write an equation expressing your evaluation.” Though this is longer, the use of the equal sign as a prompt meaning “compute” may obscure its true meaning.

use of the equal sign is critical to developing sound mathematical reasoning skills, and teachers can add to their students' understanding by modeling its usage and encouraging it. Teachers may also want to model and ask students to use the multi-line format for writing equations that is illustrated above, as it can exhibit and help structure students' mathematical reasoning.

Equal signs are used in contexts other than numbers. For example, students may encounter the equal sign when working with sets:

$$\text{If } S = \{1, 2\} \text{ and } T = \{3, 4\}, \text{ then } S \cup T = ____.$$

The third equal sign is familiar; it asks for a description of the set $S \cup T$, such as $S \cup T = \{1, 2, 3, 4\}$.² Similarly to an equation involving numbers, an equation involving sets is an assertion that two mathematical objects are actually the same; it may be either true or false. An equation involving sets is true if the sets on the two sides of the equal sign have exactly the same elements, and otherwise it is false. With the notation as above, if $S_1 = \{1, 3\}$ and $T_1 = \{2, 4\}$, then the equation $S \cup T = S_1 \cup T_1$ is true (both are the set $\{1, 2, 3, 4\}$), but the equation $S \cup S_1 = T \cup T_1$ is false.

The first two equal signs in the problem above also indicate equality, but they do so in a new context: they serve to *define* the sets S and T . Since we did not know what these sets were before, when we see the phrase “If $S = \{1, 2\}$ ” we are being asked to reason from the starting point that S is now the set consisting of 1 and 2. Making mathematical sense out of the above problem requires a careful attention to mathematical syntax.

Students next encounter equations involving one or more variables. They learn that $A = b \times h$ is the formula for the area of a rectangle. Once again, two quantities are equal—the area of a rectangle is equal to its base times its height. This equation is *always* true. It is really a mathematical statement: *if* a rectangle has base b and height h , *then* its area is given by $A = b \times h$. Because of this, this equation has more flexibility than a numerical equation; if one knows two of the base, height and area of a rectangle, then one can deduce the third from it. Similarly, a geometric object of base 20, height 10 and area 150 can not be a rectangle. Just as above, this use of the equal sign may be mixed with the use of the equal sign to define notation: *if* $b = 5$ and $h = 7$, *then* $A = 35$.

Students also encounter equations in which one variable is not known. They learn to solve linear equations such as $2x + 3 = 5$. Here it is not *always* true that the quantity $2x + 3$ is equal to the quantity 5. Rather, we write $2x + 3 = 5$ to say that we are considering the specific number x that makes this true. If x does not make it true, then it is not the one we are interested in. Thus though $A = b \times h$ and $2x + 3 = 5$ are both equations, they have very different meanings. In the first case, the equation is true in general. In the second case, the equation is false in general, and we mean to say that we are only interested in the one specific value where it is true, the value that *satisfies* the equation. (Some equations may be satisfied by more than one value of the variable. For example $x^2 = 1$ is satisfied when $x = 1$ and when $x = -1$.) We can also say that the equation $2x + 3 = 5$ is *false* when

²And just as with the problem $5 + 3 = ____$, this problem could be more carefully worded so as to avoid the use of the equal sign as a prompt.

$x = 2$, for example, or more generally any value other than 1. By contrast, $A = b \times h$ is never false provided that A is the area of a rectangle of base b and height h .

Students sometimes learn to manipulate equations without understanding why they may do so, and what they are doing. *If $2x + 3 = 5$, then $(2x + 3) - 3 = 5 - 3$.* (If two quantities are equal, subtracting 3 from each side gives two different quantities, but ones that are also equal.) Thus, *if $2x + 3 = 5$, then $2x = 2$.* And *if $2x = 2$, then $x = 1$.* Because when we write the equation $2x + 3 = 5$ we are limiting ourselves to a specific value of x , students sometimes are not sensitive that the use of the equal sign to designate two equal quantities is still at work. For example it is not uncommon for a student to write $2x - 3 = 5 = 2x = 2 = x = 1$. It is vitally important that teachers work with such students so that they do not combine different, incompatible and unequal, equations.

Because the *if, then* in the above paragraph are usually suppressed and the mathematical meaning of equations changes based on context (from always true to true only for specific values), these points must be explained by the teacher. These issues also arise in solving two equation simultaneously. Asking students to solve $2x + 3y = 6$ and $-x - 1.5y = 2$ in the standard way (multiply the second equation by 2 and add) leads to the equation $0 = 10$. This equation is of course false. What we mean by this is that *if x, y are two numbers that simultaneously satisfy $2x + 3y = 6$ and $-x - 1.5y = 2$, then it must also be true that $0 = 10$.* Since it is not, there were no numbers x, y in the first place. Similarly, attempting to solve the simultaneous equations $2x + 3y = 6$ and $-x - 1.5y = -3$ leads to $0 = 0$. This is always true, however, and we can not derive anything directly from this. Instead, we notice that *if we take any solution (x_0, y_0) to the second equation (there are infinitely many), then it is also a solution to the first equation, as can be checked by multiplying the second equation by -2 .* Thus there are infinitely many simultaneous solutions to the equations; but without the *if, then* expressed the reason for this is unlikely to be clear to students.

2. Functions

“Function” is a critical concept, one that students meet many times in their education. A function is often defined as a “rule” that assigns to each permissible value (each value in the *domain*) a unique value of the *range*. This is a fairly abstract definition, and it is not easy for students to appreciate it. We meet examples early on; for one, the area of a rectangle is a function of its base and height, in fact, exactly the function $A(b, h) = b \times h$. Here the (b, h) indicates that we are to think of the b and h as being allowed to vary, but given a specific pair, we then get only one answer for A . Going from $A = b \times h$ to $A(b, h) = b \times h$ is already a substantial conceptual step.

A “rule” is a bit intangible, and a precise definition is this. A function from a set X to a set Y is a set L (possibly an infinite set) of ordered pairs (x, y) with x an element of X and y an element of Y , with the properties that (i) for any x in X , there is some y with (x, y) in L and (ii) if (x, y) and (x, y') are both in L , then $y = y'$. The “rule” says given any specific x_1 in X , look for the pair (x_1, y_1) in L (there is exactly one such pair, by (i) and (ii)). Then the function sends x_1 to y_1 . That is, L is the set that tells us where to send each point in the domain. Notice that in this exact definition, one can only talk about a

function when one knows the set X that it is defined on, that is, knows its domain, as well as the set Y of possible values that it may take on³.

It's hardly practical to generate infinite sets of ordered pairs, so often we describe the rule another way. For example, consider the rule that says: given a number x , send it to $1/(x-1)$. Thus 0 goes to -1 , and 3 goes to $1/2$. It is common to write this rule $f(x) = 1/(x-1)$. Some care is needed here, for we run into trouble if $x = 1$, since $1/0$ is not defined. So we must limit the domain X to all numbers other than 1. (There are actually many functions given by the equation $f(x) = 1/(x-1)$; we could limit x to be integers other than 1, all real numbers other than 1, etc. However, it is common to specify the domain and range by context.)

When we write the function $f(x) = 1/(x-1)$, we are specifying where to take any x in the domain, whatsoever. Notice that if we had a number called y in the domain, it would go to $1/(y-1)$. Thus the function described by $f(x) = 1/(x-1)$ is *the same as* the function described by $f(y) = 1/(y-1)$. Though the second formula uses a different variable, for any input whatsoever, we get the same output. This tells us that it is the same function. This is a subtle point that requires careful explanation.

More generally, we say that two functions f, g from X to Y are *equal* if $f(x) = g(x)$ for all x in X . "Equal" means that both are described by the same set—independent of any formulas for them. For example, $\sin^2 x = 1 - \cos^2 x$ since these functions take the same value for every real number x . Since we always get the same answer when we evaluate at any specific number, even if we got the answer in different ways, the functions are equal. We encountered this use of the equal sign—*always the same*—in the previous section.⁴ In trigonometry, equations such as $\sin^2 x = 1 - \cos^2 x$ are sometimes called *identities* to emphasize that the equality always holds, but in fact this is the general meaning of the equal sign in this context and does not require a special word.

Functions may be described in a number of ways. Sometimes one makes a table with first column x and second column $f(x)$. This can be thought of as a piece of the set L described above. Unfortunately, sometimes this is abused in the school curriculum, where it is implicitly assumed that all functions are very simple. So let us emphasize that in general, a few values do not tell us the entire function. A function sending 1 to 2, 2 to 4, 3 to 6 and 4 to 8 does not *have* to send 5 to 10. That is what a common function, $f(x) = 2x$, does, but there are lots of other functions that have these same values at 1, 2, 3 and 4 but not at 5.

An important way to describe a function is by means of its *graph*. This is particularly important in the context of real-valued functions. When students learn about graphing, they usually think of a function as a rule, perhaps given by a formula such as $f(x) = x^2 + 1$. Beginning with such a function $f(x)$, to graph it we introduce a new variable y and graph the ordered pairs (x, y) in the plane that satisfy the equation $y = f(x)$. It is the equation which provides the *condition* to tell us which points in the plane to include in the graph

³The *range* is the subset of Y that consists of the values actually taken on, that is, the set consisting of the second components of the ordered pairs in L .

⁴It is common but sloppy practice to write equality if two functions are equal at the points where their domains overlap. For example, $\log(x^2) = 2 \log x$ is true, but only for $x > 0$, while the function $\log(x^2)$ is defined for all $x \neq 0$.

(the ones that make it true) and the which ones not to (those for which the equation is false). We met an equation being used as a condition in the previous section as well. Note that the value of $f(x)$ is a number, but the graph $y = f(x)$ consists of pairs of numbers. In fact, the graph consists of exactly the pairs in our set L above; it is the plot of all points in L .

Since the condition $y = f(x)$ specifies the set of ordered pairs (x, y) that defines the function $f(x)$, we sometimes talk about “the function $y = f(x)$.” For example, it is common to say “A function is given by $y = x^2 + 1$.” The role of y is to indicate the image of x , so that as x changes, y does too. We can think of the mathematical object so described in different ways; which is best may depend on context. We can think of the rule that takes each x in the domain to the value $x^2 + 1$. Or we can think that we are being told the set L which describes the function, with satisfying the equation $y = x^2 + 1$ being the litmus test for belonging to L . Or we can think of the geometric plot of L (here a parabola), which allows us to visualize the graph of the function as a subset of the plane. Such visualization is key in allowing us to use functions to model curves in space or other physical situations. However, the differing ways that we have of thinking about a function adds an extra layer of subtlety to learning about and properly using this concept.

We may also work with collections of functions whose rules have some uniformity. If we ask “for what a does the function $y = ax^2 + 1$ have a zero?” we mean that each specific number a gives rise to a function $y = ax^2 + 1$. We then seek conditions on the number a such that the resulting function has a root.⁵ We could get the same set of a 's by asking: for which a does the equation $ax^2 + 1 = 0$ have a solution? We regard a here as fixed, and ask if there is a number x that makes the equation true. (If there is, the number x will depend on a .)

These same points of view persist in more variables. A function may require two inputs—an example is the area function $A(b, h)$ above. We may visualize it by looking at its graph in 3-space, namely the set of all triples (b, h, z) satisfying $z = b \times h$. One can also work with families of such functions, such as the family of planes of the form $z = ax + by + c$ with the coefficients a, b, c varying. And one can work with functions of 3 or more variables as well, though it is not possible to graph them in our 3-dimensional world.

3. Conclusion

In learning a language, we learn that the same sound can have multiple meanings, but that it is usually easy to discern which is meant by looking at the context. Similarly in mathematics, context is important. We use the equal sign to express a relationship that is always true ($2x + x = 3x$), to indicate that we are interested only in a specific quantity for which a relationship is true ($2x + x = 3$), and to specify a quantity in a mathematical statement (if $x = 1$, then $2x + x = 3$). When we use the equal sign in the context of functions, we are describing an abstract and in fact rather sophisticated object. Two functions with the same domain and range are equal if they give the same result for *every* point in the domain; $\sin^2 x = 1 - \cos^2 x$. When we graph a function $f(x)$ by introducing the equation

⁵In the context of most of high school, a real root.

$y = f(x)$, we may think of the equal sign in this equation as specifying the condition required for an ordered pair (x, y) to belong to the graph. In the abstract definition of function, a function is given by a set of ordered pairs, and we write $y = f(x)$ exactly when the ordered pair (x, y) is in this set.