Being Fooled by Evidence

Frank B Murray
University of Delaware and Chair of the MACTE Board of Directors

"The first principle is that you must not fool yourself—and you are the easiest person to fool." Richard Feynman, 1974 Commencement Address

One of the great strengths of the Montessori framework is that allows its adherents to confirm the Montessori system through daily classroom observations because something that *Maria* saw, predicted, or expected is always plainly visible. This comforting strength, however, carries with it a crippling weakness that plagues all adherents to any particular educational method, philosophy or world-view. This weakness is the so-called *confirmation bias*.

This bias, once understood, can be managed. But unmanaged, it misleads us. It drives us to confirm what we believe, to seek out only examples that fit our beliefs, to bolster our beliefs through finding instances, perhaps nearly everywhere, that make sense in our theory – in this case events that support the Montessori pedagogical principles.

Consider the following example of how easy it is to fool ourselves: You are asked to discover the rule or principle that generated the following sequence of three numbers – "2, 4, 6," and your inquiry permits you to generate any other three digit sequence to see if it conforms to the rule you are seeking to discover. So, your likely hunch or theory is that the sequence increases by two, and to check your "theory," you might propose, "8, 10, 12" as an instance. You would be happy to learn that that this sequence does fit the rule. To check yourself further, you might propose "14, 16, 18" or "7, 9, 11" or "5, 7, 9," etc. and learn that these sequences also fit the rule. At some point, you may believe you have enough evidence to confirm your hunch and you announce that the rule is an increasing sequence of three numbers by two.

But, you only sought to confirm your hunch. You exhibited the bias to only confirm, and in this case you got the rule wrong and were fooled. The rule in fact was simply "any increasing sequence of three numbers," which you might have discovered if you had resisted the bias to only confirm and proposed sequences that violated your hunch, such as "1, 2, 3" or "14, 19, 30," and so forth. In each case you would have learned that these sequences also fit the rule and that your initial hunch needed adjustment.

Simply seeking examples that violate, contradict, or disconfirm your hunch or theory is an effective way to manage or counter the confirmation bias. You, of course, would hope you do not find these disconfirming examples, but to avoid the mistakes the bias leads you to make, you are well-advised to search for the examples that would mean you were wrong – and hope, of course, that you don't find any.

Consider another example to see how challenging overcoming the bias can be. Let's say you are to find the rule that generates the following words, *can, fan, man,* and *pan.* Believing the rule is any three letter English word with *a* in the middle and the consonant, *n,* at the end, you check whether *ran, tan,* and *van* conform to the rule and learn that they do. But having just learned about the confirmation bias, you check whether *bat, cat, hat, mat, pat, sat* also fit the rule, and finding that they do, you probe further with *bad, fad, lad, pad, sad,* and *fax, lax, sax, tax* and *wax* and learn that they also fit the rule. Let's say also that you even tried other vowels in the middle position and found that they also adhered to the rule, so that you were all the more confident that you have discovered the rule, and assert that it is any three-letter word with a vowel in the middle.

But, you probably didn't check whether any non-words fit the rule. The rule in fact was any CVC (consonant, vowel, consonant) trigram.[1] The challenge in overcoming the harmful consequences of the confirmation bias is to take one further step in your inquiry and attempt to falsify whatever your final conclusion is. This, of course, means that your inquiry is never completed and your conclusions are always provisional awaiting additional evidence. In other words, our theories are never proven or certain. They only have standing and a place in our beliefs because no one has disproven or falsified them conclusively – so far.

The Benefits of Falsification

We have all learned in high school geometry that parallel lines do not meet no matter how far extended. It turns out that Euclid and many subsequent mathematicians could not prove this "fact." Euclid was reduced to simply postulating it -- Euclid's fifth postulate. In other words, he simply assumed it and acted as if it were true. One method of proof in mathematics is to assume the opposite of what you want to prove and show that it since leads to something so absurd and so clearly wrong (the *reduction ad absurdum* method), that your assumption must be correct since its opposite is so clearly incorrect. However, when mathematicians assumed the opposite of the fifth postulate, for example that parallel lines met at infinity, they did not find an absurdity, but rather a new "non-Euclidean" system of geometry. When they assumed that parallel lines curved away from each other, they formulated still another "non-Euclidean" geometrical system. The point here is that while your inquiry should at some point balance your confirmations with attempts to falsify, you efforts to falsify could uncover more powerful ideas than your original supposition.

Some Encouraging Examples from Child Development Research

On the whole when researchers in child development assert that the young child, owing to a certain stage of development, can't do something, other researchers, who resist the

[1] Psychologists used CVC trigrams to study human learning and forgetting because they were nonsense that could not have been learned earlier. This way they thought they were studying pure learning and forgetting because the process was uncontaminated by the person's prior experiences.

confirmation bias, have examined other circumstances of child behavior and found beneficial outcomes. The field, for example, had more or less coalesced some years ago on the view that the young child (under 7 years) was unable to take a point of view other than his own (ego-centrism). There was ample evidence that the child's behavior fit the "ego-centrism" rule or explanation. However, it also turned out that investigators (usually teachers and mothers) could find instances where the young child could select, for example, "age appropriate" toys for younger children, explain things to them in simpler language, neither of which they could do if they had not been able take the point of view of the younger child. In the end it seemed that while the young child shows ego-centrism in many situations, it is not because they are not competent to take a point of view other than their own in other situations, contrary to a prevailing absolutist stage theory.

Researchers, to take another example, of the child's ability to form a class, to classify sets of objects (red squares, green triangle, yellow circles, etc.), often tell the child to "put together the ones that go together." When a colleague of mine found a child who was simply making random groupings of the objects, she concluded the child was unable to classify. However, when she told the child to put the objects back in the box for the next child in the experiment, she was surprised to see the child picked up all the red squares, all the green triangles and all the yellow circles before placing them in the box. Had she not, in this case by chance, posed a different instruction, she would have mistakenly concluded the child did not have the capacity that her informal method revealed the child to have.

Along these lines, another colleague of mine was advised that her child should repeat pre-school because he had not mastered some cutting and pasting exercises which the pre-school faculty believed were pre-requisites for reading. However, since the child was already reading, this example shows the risks involved in assuming there is a linear order of development or only one pathway, which the preschool faculty could have discovered had they looked to see if there were readers who could not perform their preschool exercises, or if there children struggling to read who had succeeded on the exercises. On the whole, there are very few established pre-requisites in cognitive development that fit a simple rule.

In my own research I once placed two equal length sticks in a Muller-Lyer illusion configuration so that one looked longer than the other. When I asked a child why he thought one now looked longer than the other, he replied that "his mother had put it in the oven." At one level this response was a nonsensical fabrication as there was no mother or oven around. Rather than having some psychotic root, however, the response might reveal a primitive logic based on the child's observation that some things, like bread dough, come out larger from the oven than they went in. Developmental psychologists refer to these bizarre responses as "justifications at any price" that require further probing to determine what they truly reveal about the child's thinking. In general they turn out to reveal that the child was more rational than bizarre.

In another line of research, we discovered that the child's incorrect notion that the weight of a clay ball changed when the ball was flattened was more than a failure in the child's reasoning. Children, despite their error, had a coherent and logical system for their notions of an object's weight – making a clay ball harder, rougher, bigger-looking, flatter made it heavier, while making it softer, smoother, smaller-looking, rounder made it lighter.

A Word About Standardized Tests

One problem with standardized tests of curriculum knowledge, unlike teacher-made tests of content, is that they are constructed before and independently of what the teacher actually taught. Thus, they can only be loosely about what was taught. As a result they often tell us more about what the student doesn't know than what he or she does know. Again, and unlike teacher-made tests, they usually satisfy some standards for psychometric soundness, but at a price of being about something that may be of little relevance for the classroom. As an aside, when researchers have interviewed students about their performance on state-mandated curriculum tests, they find that some who got the test item wrong actually understood the tested concept, and others who got the item right did not understand the concept at all. This is one more example of the need to resist the confirmation bias that the student's performance confirms what students do and don't know. One needs to check and probe other hypotheses about what was truly behind the student's performance.

A Brief "Course" on Reliability and Validity

Consider the following example: suppose you wanted to know a person's height, but could not measure it directly. This is the typical problem – we want to know how much of *X* is understood or possessed, but cannot see directly inside the child's mind to find out. So, we measure something else, hopefully related, that we think will tell us what is in the mind.

For the sake of argument, let's say we are blocked from measuring anyone's height, but we can measure his or her arm-span, and we use that as the measure of height.[2] This is a reliable measure because on repeated measurements of arm-span we surely would get more or less the same result. Reliability simply refers to the degree of error or noise in our measurements and with reasonable care we could measure arm-span with acceptable accuracy.

The larger question, of course, is whether arm-span is a valid measure, or a good measure, of height. What kind of evidence would bolster our confidence that is was a good measure – given that for the purposes of this example we can't measure height

[2] It turns out that arm-span and height are highly correlated so each can be used as a measure of the other with minimal error. In this example, of course, we can't know this because we are blocked from getting the height measure.

directly. There are several categories of evidence upon which we might rely and upon which measurement researchers rely.

We can see if the measurement is actually about the content we want to assess. In this case, we are at least measuring something about the body that changes over time. Teachers test the content they actually taught and their test is valid on that account to the degree that the taught content is tested and no other content creeps in. (This is called content validity)

Surely we have some theory or scholarship about what we are measuring and we can see if our measure conforms to our scholarship. In this case we know height changes over time, growing quickly initially, slowing down, becoming stable for a long time, and declining slightly at the end of life. If arm-span showed the same growth pattern we have some support for our arm-span measure as a measure of height. (This is called construct validity because it is about the ideas we have constructed about height)

Our arm-span measure might predict later behavior that is associated with height -- like basketball playing, being able to stand in the deeper water of a pool, satisfying height requirements for the military, ride roller coasters, etc. (This is call predictive or criterion validity)

Our arm-span measure might be associated current features of height -- like length of stride, how high we can reach, size of clothing needed, etc. Remember, in this example, we have nothing else to go on in estimating height but arm-span. (This is called concurrent validity)

If we were to make some decision based on arm-span, we would check to see if we were right later on. If we denied or gave some opportunity to a child based on current arm-span length, we would check to see if we were right later about whether their later height was in fact adequate or not for some task or role. (This is called consequential validity)

If we had positive results indicated above from our inquiries about the associates of arm-span, we would feel confident that our measure had an adequate degree of validity and thus could be employed in “height” assessment and policy. We tend to think IQ tests are valid because their results are associated with accepted features of intellectual capacity – like further education, patent holding, higher income, accomplished performance in the professions, privileged life-styles, etc., – all the ways in which the smarter thrive. Absent any other information, in other words, the IQ test results permit the identification of persons who tend to be associated with the known features of intellect.

While the arm-span/height examples above seem far-fetched, they accurately portray the dilemma of our needing to rely on proxy measures[3] in educational assessment. The reliability of MEFS, a topic of interest in this symposium, is simply whether the same MEFS score is found on repeated administrations. Its validity is determined by whether what it asks the child to do is anything like executive functioning (content validity), whether MEFS performance conforms to behavior patterns researchers find for executive functioning (construct validity), whether the MEFS score is associated with current or future behavior that exhibits executive functioning (concurrent and criterion validity), whether, based on a current MEFS score, the school mandated additional instruction, for example, that proved to be unnecessary or unneeded, or whether the school denied admission to children who would have done well, or accepted children who did poorly, and so on (consequential validity).

In Sum -- The Better Predictor

The best predictor of future behavior is past behavior in the same, or similar situation (known as Gurthrie's law). If we wanted to predict freshman college grades, the high school grade point average is a better predictor than the SAT or ACT, for example because the freshman grade point average is based on behavior more like the behavior in high school courses than it is on the hours of the standardized test session.

So, it is the whole of the Montessori experience that affords the better predictor of the goals of a Montessori education.

[3] Authentic assessment, as it is called, provides an exception because in these cases a direct assessment can be made. If we want to know if a person can swim, play the piano, type, speak a foreign language, etc., the skill can be directly exhibited with minimal threats to validity.