

It's All About Validity

Martin A. Kozloff

When the word "validity" is applied to research, it means that:

1. *The data measure what they are supposed to measure.* For example, the number of words that a student reads correctly out of a list of 50 words is one valid measure of *reading accuracy*, but correctly naming the parts of a book is NOT be a valid measure of accuracy. It has nothing to DO with the idea of accuracy.

2. *The research findings match the facts.* For example,
 - a. If the findings "say" that 95% of the class correctly solved 80% of the math problems on a test, and if that is what DID happen, then that finding is valid. But if only 65% of the class REALLY correctly solved 80% of the math problems, then the finding does not match the facts, and is invalid.

 - b. If the findings show that students as a group made a lot more progress using a new math program than students in another class who used the old math program, and if the new math program REALLY made the difference, then the finding is valid. But if students who got the new math program learned more NOT because of the new program but because their teacher added practice, or because the experimental (new math) class was brighter than the class using the old program, or because some children's families helped them learn math, then the finding that the new math PROGRAM made the difference is INVALID.

 - c. If the group (**sample**) of students and schools with whom research was done is small or is different from the **larger population** of students and schools, the findings from the research on the "unrepresentative" sample may NOT

APPLY (not generalize) to OTHER students and schools. If researchers say that the findings DO apply to the larger population, their statement is INVALID.

Let's look at features of research that affect validity. This will help you to tell the difference between research you can trust to make instructional decisions and research that is questionable; for example, decisions about which programs and methods are effective and which are not.

Imagine that a member of your family is sick. You go to a physician. The physician "runs some tests" and comes back with a diagnosis.

"Your brother has a blood infection. It is very serious. I'm going to put him on a new drug, Viracide."

Are you going to say, "Okay, Doc. Whatever you say is fine with me," or do you have some questions?

Of course, you have questions. For example

1. What tests did the physician run?

Did they DIRECTLY show blood infection or was the doctor guessing?

How reliable are the tests? For example, if different lab technicians ran the same test on the same blood samples, would they get the same results? If not, then are the results correct (valid)? If not, maybe your brother has something else that's wrong. Or nothing wrong.

2. The physician concluded that your brother had a blood infection throughout his body. But did the physician draw blood samples from all

over your brother's body, or just from one finger? If just from one finger, then maybe the test results (infection) really apply ONLY to the finger.

3. The physician prescribes a new drug, Viracide.

Was Viracide tested on persons who had the same illness as your brother?

If not, then why does the doctor think it will work?

Was Viracide tested on many persons? Thousands? No? Then why does your doctor think it will work with your brother. Maybe it only works with *some* of the persons who were in the small sample on which it was tested?

Was Viracide tested against no treatment and/or other treatments? In other words, were there *experimental groups* that got Viracide and *control groups* that did not? If there were no control groups, then how does the doctor know that Viracide did anything. Maybe persons in a control group that got no treatment would *also* have improved. And maybe persons who got a different treatment would have improved as much or more than persons who got Viracide.

If experimental and control groups were used, were the participants *pretty much the same* in all groups, so that they only BIG difference was treatment? If the experimental and control groups were *different in several ways* in addition to being different in treatment (one group got Viracide and the other did not), then these differences (and not Viracide) may account for improvement. For example, some members of the experimental (Viracide) groups may have had healthier diets, or had more social support, or were healthier overall. If so, then maybe THESE differences, and NOT Viracide, are why the experimental groups improved more than the control groups. Or, maybe members of the control groups had healthier diets, more social support, or were healthier overall. If so, the control groups might have improved as much as the Viracide groups---

not because Viracide doesn't work, but because these OTHER factors also help persons get well.

Was Viracide tested again and again (*replicated*)? If not, then why does your doctor think that it actually works? Maybe the persons who got it and improved would have improved anyway, without it.

Was Viracide tested on many samples of persons who were different in many ways—for example, age, seriousness of the infection, general health? If not, then why does your doctor think it will work with your brother, who may NOT be like the sample of persons with whom it seemed to work?

When Viracide was tested, what kinds of data were taken to check if Viracide was working? Were there *DIRECT quantitative measures* of infection (for example counting the amount of virus in blood samples)—so that researchers could SEE if it was working? Or was the main measure merely the *verbal subjective qualitative opinion* of patients? “Oh, yes, I feel much better.”

Do you want your brother to take a medication whose only test was whether patients said they *felt* better?

When Viracide was tested, how long did the tests last? Were the studies *longitudinal*? If they were of short duration, then how does the doctor know whether Viracide has long term good effects or if it only worked for a few months and then persons got sick again?

When Viracide was tested, was the amount of blood infection measured before Viracide was given, every few weeks while it was given, and at the end of treatment (a *pre-test/post-test design*)? If blood infection was

measured only at the end of treatment (a *post-test-only design*), then how does the doctor know that Viracide did ANYTHING? Maybe the levels of blood infection were low not only *after* treatment but BEFORE treatment as well. In other words, unless you compare the amount of infection before and after treatment, you can't tell if Viracide has ANY effect on blood infection.

Now let's apply these questions about medicine to education. What do we want to see?

Let's say the reading achievement in a school, or district, or state is too low. Students are given reading tests. The test results show that reading achievement is so low that the problem is serious. Administrators decide that different reading programs are needed. Are the data, the findings, and the decision to use a new program valid? Let's see.

1. How was reading evaluated? Did the reading tests DIRECTLY measure reading skills or did they measure something else? For example, did students get points if they guessed at what words said, or did they get points off if they guessed rather than sounded out the words? Do you WANT guessing to be considered reading?

Were all five reading skills measured (see the materials on Reading First): phonemic awareness, alphabetic principle (knowing the sounds that go with letters; using this knowledge to read words), fluency, vocabulary, comprehension? If fluency, vocabulary, and comprehension were NOT measured, then the tests did NOT give a valid picture of how well students read. Therefore, how can decisions be made about how to improve reading? It would be like prescribing a medicine when you don't know what is wrong.

Were the tests quantitative and objective (for example, did they measure the number of correct words students read per minute), or were the tests and data subjective and qualitative? “My students read pretty quickly.”

If the data are subjective and qualitative, then how do you know they are accurate? And how do you know what they MEAN? What is “pretty quickly”?

2. Administrators concluded that reading achievement was a serious problem. But did they examine test data on different *subgroups* (e.g., ethnic groups, social class, sex, grade level) or did they only use test data for schools *as a whole* (e.g., 65% of students pass the tests)? If they don't have data for different subgroups, then how do they know WHOSE reading is better and whose is worse? Maybe disadvantaged children read much more poorly than affluent children. If so, are you going to give these children the SAME new reading programs? That would be like using the same medicine and the same dose regardless of the different illnesses patients have.

3. Administrators select a new reading program. Who says it was a valid selection?
Were the features of the new program (let's call it Reading Blast) based on scientific research? For example, are the different reading skills taught in a sequence that has been shown to be effective? If not, then Reading Blast is like a fake medicine---“Doctor Feelgood's Dynamic Elixir and Nerve Tonic.”

Was *the whole Reading Blast program tested* in real schools? No? Than who says it works? Maybe it is well-designed. But maybe it is so hard to use that teachers do a poor job with it. Imagine a medicine that works but it is in a package that is so complicated you can't open it.

Was Reading Blast tested with students, teachers, and schools like yours?
No? Then who says it will work with yours?

Was Reading Blast tested on many students with many teachers in many schools? No? Then maybe when it SEEMED to work, it was just a fluke. For example, you have a recipe for tuna casserole that YOUR family (five persons) loves. But try it with the whole neighborhood! Half the neighborhood might throw it in the garbage.

Was Reading Blast tested against other reading programs---including the old one that your school used? In other words, were there experimental groups that got Reading Blast and control groups that did not? If there were no comparison groups, then how do you know that Reading Blast works better than what you used to use or works better than other programs? Why invest children's time and teachers' efforts on something that may not be any better?

If experimental and control groups were used, were the participants pretty much the same in all groups, so that the only BIG difference was Reading Blast? No? Well, what if the students who received Reading Blast in the tests already read a little, or had inventive teachers, or had parents who were teaching them to read at home? If so, it will look as if Reading Blast is very effective when in fact OTHER factors are helping the students to read. Therefore, Reading Blast won't do much in your school because your children do not have the OTHER factors going for them.

Was Reading Blast tested on many samples of students, teachers, and schools that were different in many ways—for example, beginning reading skills, teacher proficiency, ongoing assistance to teachers, family

involvement? No? Then who do you know Reading Blast will work in your school?

When Reading Blast was tested, what kinds of data were taken to check if it was working? Were there DIRECT quantitative measures (for example counting the number of correct words children read from a list)—so that researchers could SEE if it was working? Or was the main measure merely the verbal subjective qualitative opinion of teachers? “Oh, yes, they are reading much more accurately with Reading Blast.”

When Reading Blast was tested, how long did the tests last? Were the studies longitudinal? If they were of short duration, then how do you know whether Reading Blast has long term good effects or if it only works for a few months and then students’ progress slows way down, or teachers get real bored?

When Reading Blast was tested, were reading skills measured before Reading Blast was started, every few weeks while it was being used, and at the end of the semester and year(a pre-test/post-test design)? No? Well, without a pre Reading Blast, during Reading Blast, and post Reading Blast comparison of reading skill, how can you tell if children learned ANYTHING? Or when they learned it?

How many different measures were used to test each reading skill when Reading Blast was used? For example, how many ways did researchers check to see if students could sound out words? Did they use only one way: for example, students read words from a list? But that is not the same as students reading words in sentences (connected text). Maybe students who received Reading Blast DO become more skilled at reading words from a list but NOT more skilled at reading connected text. In this case, is Reading Blast going to be much good in the long run?