Consequential Aspects of the Validity of Achievement Tests:
A Publisher's Point of View

by

Donald Ross Green

CTB/McGraw-Hill

Submitted to Educational Measurement: Issues and Practice
January 21, 1998
The principal thesis of this paper is that the publishers of achievement tests, particularly the publishers of those tests intended for use in many districts across the nation, are, for the most part, not in a position to obtain on their own any decent evidence about the consequences of the uses made of their tests. Some of the reasons why this is true are offered, followed by examples of consequences that have been claimed and/or investigated, leading to the conclusion that there is little credible and/or useful hard evidence on this matter. Finally a brief attempt is made to specify what actions publishers of such tests can reasonably be expected to take with respect to the many and varied consequences of use. Also discussed are the kinds of help that need to be provided by the other parties in the enterprise if any solid information on the issue is to be obtained.

Factors that Limit Publishers' Abilities to Examine Consequences

The reasons why publishers of achievement tests are not typically able to investigate this aspect of validity are readily seen.

The tests discussed here are the familiar and widely used nationally normed and nationally marketed achievement batteries. Typically these tests are designed, developed, and normed over a three- or four-year period, and substantial use of them often is not made until some five or more years have passed from their conceptualization.

The uses that are made of such tests are numerous and vary by teacher, by school, by district, by state, and over time.

General discussions of their validity among school personnel from different districts can be confusing because these batteries are collections of up to one hundred or more different tests; no school uses all of them and almost no schools use only one of them. Some of the tests within
the batteries are very different from each other and generalizations across the set are often ill-founded. This problem is exacerbated when the batteries in question are not the same. Unfortunately some people believe that all such tests are alike, even though a conscientious review of these tests would surely disabuse the reviewer of that conviction.

The responsibility for clear and convincing descriptions of the constructs being measured is obviously the responsibility of the publisher, but any faith that such statements have much to do with interpretations and uses of the scores is naive. Regardless of whether the constructs have been clearly described by the publisher, each teacher, curriculum coordinator, test director, superintendent, school board member, state department official, state legislator, news reporter, and member of the various advisory and review committees has his or her own view of what constitutes reading, mathematics, science, social studies, and so forth.

No direct mechanism exists for obtaining really credible evidence of the many possible consequences of the use of these achievement tests. Few if any schools or districts collect such evidence in a scientific manner. Typically test scores are only one source of information among many that affect instruction. Furthermore, the typical school system uses a particular achievement test for about four or five years and then changes to a new test; therefore, by the time evidence of the consequences of their use of the tests could have been accumulated, the school system is no longer interested in that test. Establishing long-term experimental and control groups needed for separating the effects of test use from the effects of other correlated conditions and instructional practices appears nearly impossible.
Existing Evidence About the Consequences of Standardized Test Use

Some reliable evidence does exist about the consequences of using standardized achievement tests. Probably every publisher has some but I will limit myself to a few examples from CTB’s experience. The “consequential aspects of validity” may be relatively new terminology but concern about this matter is not new.

Criterion-referenced tests. The very first thing I did when I came to CTB in 1967 was to ask what uses were made of our tests, the California Achievement Tests in particular. The answer was the same then as it would be now from most people at CTB, to wit: The leading purpose of these achievement tests is “to help the teacher help the child.” When asked how that was accomplished nobody seemed to really know. Therefore, I went to a nearby school district and met individually with about ten elementary teachers in various grades and asked them what they did with the results. That particular investigation stopped there because none of them could name any concrete action they had taken from the test score data other than using it to talk to parents, and only a few of them did that. I soon learned that many teachers did in fact use the results of the “Diagnosis of Learning Difficulties” based on a report showing right/wrong for each item. Many of us at CTB were bothered by the unreliability of what were often single item scores; the ultimate upshot of all this was to attempt to make test scores more useful to teachers and students by moving into criterion-referenced tests, which we began to publish in 1970.

Throughout much of the 1970s and early 1980s we conducted “learner validation” studies for these criterion-referenced tests. These studies collected evidence about what happened when teachers used the results of these tests on an individual basis. The studies had many procedural and operational problems, but generally they appeared to show that student achievement on tests
specific to the objectives taught exhibited sharp gains in score (CTB/McGraw-Hill, 1984; Gessel, 1985). Did these programs help students? I truly believe they did but teachers found them complicated to implement, and in the early 1980s the interest in specific objectives began to fade. Consequently the necessary long-term follow-up studies were never carried out and these sort of criterion-referenced tests disappeared much like their predecessors of the late 1920s.

Do traditional norm-referenced tests narrow the curriculum? Consider the claims that traditional norm-referenced tests narrow the curriculum and that therefore students learn less when they take such tests (Frederiksen, 1984; Madaus, West, Harmon, Lomax, & Viator, 1992; Shepard & Dougherty, 1991). There is a visible likelihood that, in some instances at least, this is true based on the logic of the situation and reports of what teachers say they do (Madaus et al., 1992), but where is the experimental evidence? As far as I know, Shepard et al. (1996) made the first serious attempt to carry out such a study. Their findings do not show much support for the contention. While this result is probably for the reasons they offer, it remains to be demonstrated that students really do learn less when schools use traditional standardized tests.

Do performance assessments improve learning? Then there are the reports of the consequences of the “new sorts of assessments,” namely performance assessments. For example, Kentucky has reported sharp gains on the Kentucky Instructional Results Information System (KIRIS) and has suggested that this outcome arose because the testing program has led to better learning and instruction (Kentucky Department of Education, 1995). Perhaps the inference from the score gains is justified and I, for one, would certainly like to think so. But one cannot help noticing that a similar phenomenon, i.e., rising test scores, led to talk of “teaching to the test” and “the Lake Wobegon effect” just a few years ago when the tests in question were multiple-choice tests (Shepard, 1990; Phillips, 1990).
It is also notable that CTBS scores, which had been rising when that test battery was the official evaluation of the Kentucky districts, stopped rising. Now for some people, that merely indicates that what such multiple-choice achievement tests measure is irrelevant to “real learning.” However, if that is the case, how does one explain the following facts?

- As students go up the grades they score higher on such tests.
- Generally acknowledged “good” students almost always score much higher on such tests than those not so acknowledged.
- Teachers in the content area and grade as well as content experts almost always score even better than good students on these multiple-choice tests.
- Correlations with performance assessment scores are usually substantial given the limitations in reliability of both measures.

One interpretation of these results is that the students in Kentucky, while maintaining their scores on CTBS, have not been able to generalize the greater knowledge and skill exhibited by the increase in scores on KIRIS. Obviously there are other possible interpretations. For example, because only some districts chose to give CTBS in those years and because their uses varied, both teacher concern and student motivation to perform may have been less uniformly focused than was the case with KIRIS.

A less striking, but possibly similar, result appears in Maryland, where the CTBS statewide scores clearly stopped rising when it was no longer the state test. The Maryland School Performance Assessment Program scores in reading went up the first year but not the second, while in mathematics, somewhat lesser growth the first year was followed by a little further growth the second year. The variation from district to district in these patterns in both
tests and their relation to each other is notable. Counterexamples of almost any interpretation can be found in these data (Yen, 1996).

Thus, it appears that those who believe that performance assessment necessarily improves instruction have yet to make their case, and I believe that the data just cited opens to question the assertions about the evils of multiple-choice tests. The evidence about these “consequences” is far from solid.

*Do multiple-choice tests inhibit conceptual learning?* Another common assertion is that multiple-choice tests encourage or even require the memorization of isolated facts and inhibit problem solving and depth of conceptual learning. It is hard to believe that many teachers allow this to happen but perhaps it does happen, either because of, or in spite of, the teachers. D’Ydewalle, Swerts, & De Corte (1983) reported a study indicating that students who were told that they would be given an essay exam did better on a multiple-choice test than did students who were told that it was going to be a multiple-choice exam partly because the latter group studied longer but apparently also because they studied differently. This finding lends support to the assertion as does a recent study by Thiede (1996). On the other hand, Hakstian (1971) concluded that there was no such effect. Lundeberg and Fox (1991) point out that Hakstian did in fact report a significant difference favoring students with an essay set on multiple-choice items measuring analysis. However, Lundeberg and Fox conclude from their review of this matter that the data and research available are too thin to draw a conclusion.

The various studies and discussions of this issue suggest that students get impressions of what the tests are measuring mostly from teachers but also from each other and that these impressions do have consequences. However, note that neither of these categories of sources have ever, nor will ever, look at any publisher’s statements of the constructs being measured. In
short, this is an example how disconnected publishers are from the uses of their tests and why we cannot really respond well to the various public assertions about the consequences of the uses of our tests.

**Role of Opinion Versus Evidence**

There are, of course, many people making a variety of assertions about published tests, a few of which have just been described. Naturally publishers devote much effort to understanding these opinions; they do not operate with blind faith in evidence, rather they collect opinions. They get customer comments, they get customer complaints, they get customer questions—often about interpretations and uses. Some customers volunteer what they are doing with the tests, and sometimes they make claims about good or bad consequences; but what is really rare is solid scientific evidence. Publishers usually interview educators and conduct focus groups of customers and others. However good these sources may be, the nature of these data on consequences for students is at best second or third hand, tends to be anecdotal, and is almost always hearsay. Even the measurement community tends to rely on evidence of test consequences that is of this nature (e.g., Koretz, Stecher, Klein, & McCaffrey, 1994; Koretz, Mitchell, Barron, & Keith, 1996; Shepard, 1990).

Note that the demonstrations of the good consequences that followed from the use of criterion-referenced tests did not lead to more use of these kinds of tests. Opinions, not evidence, about what are “good” tests appear to be the far more potent element in the marketplace.

Publishers pay serious attention to assertions about the consequences that flow from the use of the tests, especially when they are made by the widely quoted currently influential academics. So for all those who believe they know or have better ideas about what tests and
testing programs should be like, this is a call to offer solid evidence about the consequences of
the changes and improvements they are touting. For now, I submit that the “value implications”
of these various score interpretations (Messick, 1989) are inadequately evaluated.

What Publishers Can Do to Examine Consequences

What then can and should publishers do to meet their responsibilities? The options in
order of increasing desirability and reasonableness are:

1. Ignore the issue and/or insist that it is entirely someone else’s responsibility.
2. Undertake to seriously study the matter by themselves for each of their instruments.
3. Try to persuade academic researchers to study the matter objectively.
4. Try to work out some cooperative studies with individual customers. Where they exist, try to
   persuade the Technical Advisory Committees to design and recommend such studies.
5. Work through organizations such as NCME to get a series of systematic studies of the matter
designed, financed, and staffed involving many publishers, many school systems, and many
academics.

The merits and problems with most of these options are numerous so only a few will be noted.

The first option is clearly unacceptable, even though it comes uncomfortably close to
representing the status quo. The second and third options should perhaps be encouraged. A few
serious studies might eventually appear in technical reports and a few others in journals some
years later. However, the relevance of these reports to the testing programs then being set in
place will almost always require generalizations well beyond the reported data and extrapolations
to situations which differ in ways whose significance for the inferences is unknown. In any case,
I am sure everyone recognizes that the difficulties and limitations of drawing inferences from one or a few isolated studies are legion. In this instance one would wish to ask:

- Which users? There are in the neighborhood of fifteen thousand school districts that use these tests.
- Which tests? The typical battery may have up to one hundred or more different tests; they cover anywhere from five to fifteen content areas in varying formats and differ substantially in content from grade to grade.
- Which uses? While there are probably not more than ten or a dozen major uses of the scores, the variations in the way in which these are executed are quite large and no one knows which variation has which effect.

Given the number and range of such issues, it seems self-evident that only a large scale cooperative approach has any hope of shedding light on the general issue. Given that kind of cooperative effort, perhaps generalizable results might be possible for existing tests and some distinctions between the consequences of various kinds of achievement tests might be found.

Unfortunately the problems cited above for isolated studies are not necessarily solved by a cooperative effort because there will be additional problems:

- Few school systems are likely to welcome reports of unanticipated negative consequences of their testing programs, so cooperation may be hard to obtain.
- Agreement among interested parties about the appropriate criterion measures of the consequences is likely to be contentious at best.
- Any cause-effect conclusions are likely to be disputed endlessly.
• If what has happened to date in the evaluation of performance assessment is any indication, much of the research undertaken is likely to be by those trying to prove that whatever exists is inferior to their new and better idea, which of course will not be tested for many years. I am sure this list of problems is incomplete.

In short, given the circumstances that were described at the beginning of the paper, it should be apparent that a well-designed set of related studies would need to be carried out to yield substantial believable results applicable to more than a few of the uses made by a few of the customers of any one of the publishers of the nationally normed batteries.

I repeat that I believe it will take a large scale cooperative effort to produce any generalizable evidence about the consequences of the use of nationally normed tests, whatever their formats. It is probable that all the major publishers would be willing to participate in such an effort; certainly CTB would like to do so.


