Chapter 9. Cause or Effect? Validating the Use of Tests for High-Stakes Inferences in Education

Derek C. Briggs
University of Colorado at Boulder
July 2009

The author thanks Neil J.

9.1 Dorans and Brent Bridgman for helpful comments on an earlier version of the manuscript. Much of the research described in this chapter was supported by a National Academy of Education Postdoctoral Fellowship funded by the Spencer Foundation.

Formatted: Title, Left

Formatted: Author, Left

Formatted: Heading 1, Left

9.1 Introduction

"Casual ~~C~~comparisons ~~I~~inevitably ~~I~~initiate ~~C~~careless ~~C~~causal ~~C~~onclusions."
—Paul Holland, 2004

A good aphorism can, in a few words, capture an essential truth. Of the many good aphorisms Paul Holland has coined over the years, I have found myself invoking the one above frequently enough to worry that I should be paying out royalty fees, so it is only fitting that I use it as the starting point for some ideas I wish to explore in this paper.

It is fairly common for people to use the graphical shorthand $Z \rightarrow X$ to represent the inference that a change in some variable "~~Z~~" causes a change in another variable "~~X~~". Yet without further explication, this sort of presentation is causally ambiguous. In his seminal presentation of what he termed "*Rubin's ~~Causal~~ causal ~~Model~~model*" (also known as the ~~Potential~~ potential ~~Outcomes~~ outcomes ~~Model~~model~~,~~ or the Neyman-Rubin ~~Model~~model), Holland (1986) clarified the elements necessary to define and estimate a quantity interpretable as an average causal effect. These elements include the units of analysis, the specific treatments to which units may or may not be exposed, the potential outcomes as a function of treatment exposure, the mechanism by which units are exposed to treatment conditions, and the approach taken to estimate an unbiased average causal effect. In theory, the application of Rubin's ~~Causal~~ causal ~~Model~~ model for the design and analysis of an experiment or quasi-experiment should serve as a safeguard against drawing "~~~~careless" causal conclusions. However, ~~I believe~~ this safeguard has an Achilles Heel in the context of its application in educational research: the often--equivocal nature of test scores as measures of cognitive outcomes.

Rubin's ~~C~~causal ~~M~~model is agnostic about the measurement properties of the test used to define these potential outcomes: ~~T~~the role of a test score is to provide the units through which the estimate of an average causal effect can be quantified. My contention is that in many circumstances a failure to think carefully about test validity will serve to undermine inferences about an estimated average causal effect, whether or not this effect is unbiased in the statistical sense laid out by Holland (1986).

As measures, not all outcomes are created equally. For example, death and income are common outcome measures in epidemiology and economics, and are relatively straightforward to validate. In educational research, cognitive outcomes are typically of interest, but such outcomes are unobservable. Cognitive outcomes are measured with standardized tests, and the match between test scores and their intended interpretation and use has spawned a dense literature in psychometrics under the umbrella term of validity theory. In this paper, I will be making an argument that at first glance appears either circular or paradoxical: C~~c~~ausal inference in educational research depends upon establishing test validity, but test validity depends upon establishing a causal inference. The reason I ~~develop~~developed this argument is because I think it can serve the purpose of helping to bridge the gap between validity theory and practice in the context of high-stakes test use in education. That is, once we see that causal inference and test validity have a symbiotic relationship, it becomes possible to kill two birds with one

Formatted: Font: Not Bold, Not Italic

Formatted: Heading 2, Right, Line spacing: single

Formatted: Right

Formatted: Body Text, Indent: First line: 0"

Formatted: Font: Italic

stone: I~~i~~n estimating the effect or effects of educational interventions, we may also gain valuable insights about what it is that tests are (and are not) really measuring.

~~There are f~~Four sections ~~that~~ follow. In the first section, I provide a policy context for the kinds of causal inferences being made about education in the wake of ~~the~~ No Child Left Behind ~~law~~ (NCLB; ~~NCLB;~~ N~~o~~ C~~hild~~ L~~eft~~ B~~ehind~~ Act of 2001, 2002). I suggest that a focus on making causal inferences that are internally valid has overshadowed the important role played by the choice of test outcome in causal generalization. In the second section, I provide a brief overview of current conceptions in test validation theory~~,~~ and contrast this with current state-level practices. I ~~(~~re~~)~~introduce an idea dating back to at least Cronbach (1971) that test validity might be fruitfully evaluated through the lens of causal inference and experimental design. In the third section, I elaborate upon a validation design that uses the real-world context of NCLB-mandated tutoring as the basis for an evaluation of the item level instructional sensitivity of large-scale assessments. In the fourth section, I offer concluding comments.

9.1.1 The Context of Causal Inference in Educational Research

It would be difficult to overstate the impact NCLB has had upon state systems of educational accountability since its implementation in 2002. The stipulations of NCLB require ~~that~~ all schools receiving Title 1 funds to test their students annually in the subjects of math, English/language arts, and science in grades 3 through 8 and at least once during high school. The performance of students within a given school (disaggregated by demographic subgroups) is then evaluated relative to criterion-referenced thresholds for each subject-specific test. Students are subsequently classified into performance levels (~~,~~ e.g., ~~"~~unsatisfactory~~"~~, ~~"~~proficient~~"~~, ~~"~~advanced). ~~"~~By the year 2012, ~~the~~ a target ~~is~~ was set that~~for~~ 100% of students ~~to~~ should demonstrate test performance that would place them in the proficient category or higher. To this end, states were asked to specify target school-level percentages of students classified as proficient or higher each year leading to 2012. Each year, if a school's aggregate percentage is below the target percentage for any student subgroup or test subject, they will have failed to demonstrate ~~,"~~*adequate yearly progress*~~"~~ (AYP). High-stakes sanctions are attached to the NCLB law. If a school fails to make AYP in ~~two~~ 2 consecutive years, it must offer parents the opportunity to choose a different public school for their child to attend. After ~~three~~ 3 years of failing to make AYP, supplemental educational services (i.e., tutoring) must be provided for all students eligible for free or reduced lunches. After ~~five~~ 5 years of failing to make AYP, schools become candidates for restructuring by an external agency.

The extent to which NCLB has had a positive or negative impact on the American educational system is unclear. However, the law has achieved one important ancillary outcome: I~~i~~t has established a tremendous infrastructure for evaluating the causal effects of educational interventions. When NCLB was authorized in 2002, relatively few states tested grade 3 through 8 students annually in multiple subjects, and only 18 had a statewide identification system in place that could link students, their test scores, and their schools over time. ~~As o~~Five year later by ~~f~~ 2007, virtually all states we~~a~~re testing students in grades 3 through ~~-~~8 in math, English/language arts, and science, and ha~~d~~v~~e~~ a

Formatted: Heading 3, Left, Indent: First line: 0", Don't keep with next, Don't keep lines together

Formatted: Font: Italic

Formatted: Body Text, Indent: First line: 0"

statewide student identification system. Combined with the use of the I~~i~~nternet as a means of transferring large quantities of data electronically in a timely and secure manner, the upshot is the availability of longitudinal data for research and evaluation purposes on a scale previously only possible through federally funded surveys conducted by organizations such as the Department of Education's National Center for Education Statistics.

The scores from the various standardized tests being administered from state to state are now being used to facilitate a host of evaluative studies. I want to distinguish two types of prevalent studies ~~that are prevalent~~. The first type of study is an evaluation in which a specific educational intervention has been implemented; the second type is one in which pre-existing teachers and/or schools are themselves under evaluation as an educational intervention. In both cases, causal inference hinges upon the following question: W~~w~~hat is the effect of a given intervention on one or more cognitive outcomes? The answer to such a question can have high-stakes ramifications: Curricula may be adopted or abandoned; teachers may receive salary increases or get fired. Given that the causal inferences are high-stakes, it is clearly important to get the magnitude and direction of effect estimates right. But it is just as important to make sure that appropriate test scores are being used as outcome measures. I next describe two empirical examples from published studies, one for each of the study types defined above, in which the choice of outcome measure ~~can lead~~led to very different causal inferences about the effect of an educational intervention. In both these examples, I focus on the domain of mathematics proficiency in the middle schools grades, and I put to the side the issue of whether any given causal effect estimate is in fact unbiased.

Formatted: Body Text

Formatted: Heading 3, Don't keep with next, Don't keep lines together

9.1.2 Evaluating the Effects of the Connected Mathematics Curriculum

Formatted: Body Text, Indent: First line: 0"

Beginning in the late 1980s, the National Council for Teachers of Mathematics (NCTM)~~,~~ published a series of documents describing new standards for how math should be taught at different grades. The standards called for a greater emphasis on knowing when and how to use mathematical skills and concepts to solve real world problems. The Connected Mathematics Project (CMP)~~,~~ was funded by the National Science Foundation (NSF) to develop a "~~"~~reform-based~~"~~ math curriculum for grades 6 through 8, a~~. A~~s described by Ridgway et al. (2000~~, p. 182~~):

Formatted: Block quote para 1, Indent: Left: 0", Adjust space between Latin and Asian text, Adjust space between Asian text and numbers

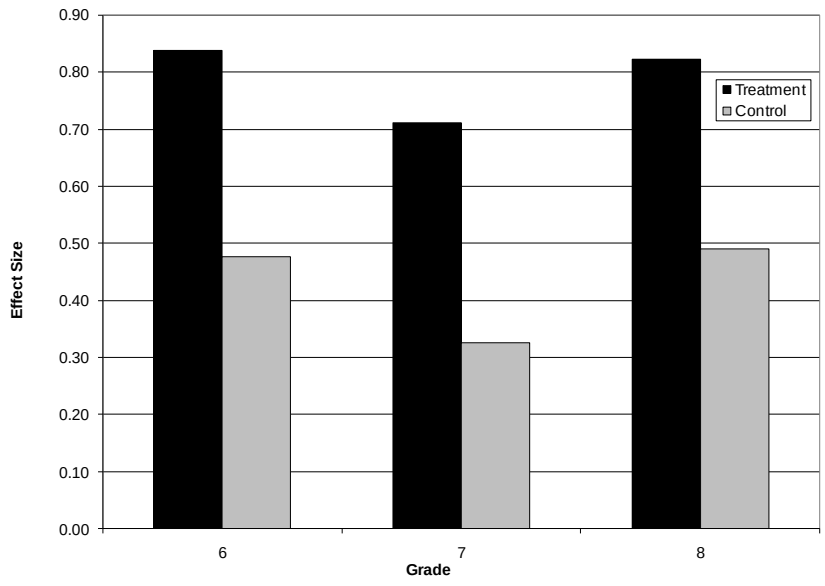
> The CMP curriculum is organised around problem settings. Activities are designed to involve groups of students with mathematical concepts and applications, and in discourse and reflective writing about these same ideas. Students are expected to observe patterns and relationships, make conjectures, discuss solutions and generalise from their findings. The goal is to immerse students in the mathematics and the styles of mathematical thinking needed for success in high school and eventually college. (p. 182)

As a means of evaluating changes in student understanding during exposure to reform-based mathematics curricula, the Balanced Assessment (BA) was developed in a concurrent project also funded by the NSF (Ridgway & Schoenfeld, 1994). According to Ridgway ~~et al.~~, Zawojewski, and Hoover (2000), the BA test was not designed such that its tasks ran in parallel with those on the CMP curriculum; rather, the aim was to assess

transfer of learning according to the educational goals set out by the NCTM Standards. The BA tests consist entirely of open-ended items designed to assess reasoning, mathematical communication, connections, and problem solving. Because the open-ended items are time-consuming to complete, only a subset ~~are~~ is administered to any given test-taker in one of five forms. Each form contains 10 to ~~-~~15 individual items ~~which~~ that are scored both holistically and analytically by trained raters.

Ridgway et al. (2000) reported on the results of a quasi-experimental evaluation of the CMP curriculum. The study employed a pre-post design with two different tests: One test was the BA described above; the other was the Iowa Test of Basic Skills (ITBS). The ITBS consists solely of multiple-choice items that focus on the mastery of technical skills in mathematics. A total of 500 ~~sixth~~ grade 6 students, 861 ~~seventh~~ grade 7 students, and 1,095 ~~eighth~~ grade 8 students took grade-specific versions of these tests at the beginning of a fall semester~~,~~ and then again at the end of a spring semester. In each grade, some students were taught math using the CMP curriculum (reform-based treatment condition), while others used~~ing~~ commercially available textbooks (non-reform-based control condition). The authors subsequently compared the standardized gains for each group as a function of the outcome measure being used. These results are presented graphically in Figures 9.1 and 9.2.

When evaluated using the BA tests, the results seem unequivocal. As shown in Figure 9.1, students exposed to the CMP curriculum have considerably larger average gains than students exposed to traditional curricula. I-n contrast, when evaluated using the ITBS, ~~there is~~ far less compelling evidence exists to support the effectiveness of the CMP curriculum. There appears to be a negative effect in grade 6, no effect in grade 7, and a positive effect in grade 8.



Formatted: Body Text, Indent: First line: 0"

Formatted: Figure Caption, Left

Figure 9.1. Standardized gGains on Balanced Assessment (BA) tTests bBy gGrade and cCondition. [Derek: Are Figure 9.1 and Figure 9.2 taken from Ridgway et al. (2000)?]

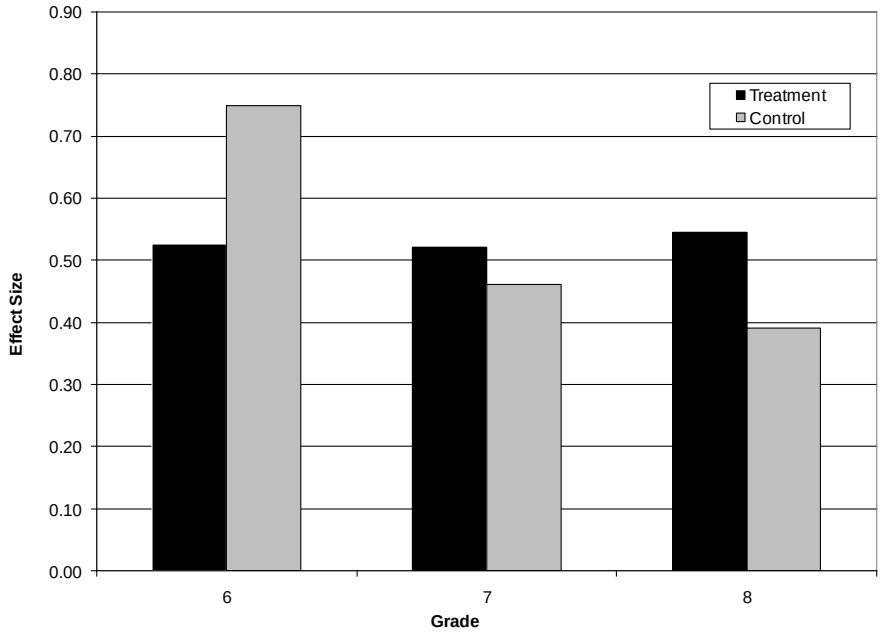


Figure 9.2. Standardized gGains on Iowa Test of Basic Skills (ITBS) tTests bBy gGrade and cCondition. [Derek: Are Figure 9.1 and Figure 9.2 taken from Ridgway et al. (2000)?]

A couple of comments are in order. First, the items on the ITBS are likely to be very similar to the types of multiple-choice items on the state-level tests administered to fulfill the requirements of NCLB. They are not necessarily bad items, nor is the ITBS necessarily an invalid test. However, the ITBS was not designed to evaluate the same cognitive outcomes for which the BA test was designed. If the ITBS ~~was~~ were used as the sole outcome measure to estimate the effect of the CMP curriculum in grade 6, one would be likely to draw the conclusion that the curriculum should be abandoned. By contrast, were the BA test to be used, we would conclude that the CMP curriculum should be celebrated. Second, the different patterns of findings by test are the kinds of results that can lead to a greater understanding of the curriculum under investigation~~,~~ and how children are learning. A typical argument by those developing curricula that supposedly focus on depth of conceptual understanding is that this will not sacrifice "surface" understandings that are more procedural. The results from the Ridgway et al. (2000) study suggested that procedural understanding (as measured by the ITBS) ~~may~~ might suffer when students have only been exposed to ~~one~~ 1 year of the program, but for students exposed to ~~three~~ 3 years of the CMP curriculum, this gap reverses.

Formatted: Highlight
Formatted: Highlight
Formatted: Highlight
Formatted: Highlight
Comment [DB1]: NO—I created these fiigures on my own from their results
Formatted: Font: Bold, (Asian) Japanese, Highlight
Formatted: Font: Bold, (Asian) Japanese, Highlight
Formatted: Font: Bold, (Asian) Japanese, Highlight
Formatted: Font: Bold, (Asian) Japanese, Highlight
Formatted: Figure Caption
Formatted: Body Text, Indent: First line: 0"
Formatted: Body Text

9.1.3 Evaluating the Effectiveness of Teachers with Value-Added Models

Formatted: Heading 3, Don't keep with next, Don't keep lines together

Formatted: Heading 3

Formatted: Body Text, Indent: First line: 0"

Value-added modeling (VAM) has become increasingly popular in the context of educational accountability systems because it offers the potential to estimate the effect of a specific teacher or school on student achievement independent of the influences of race, socioeconomic status, and other contextual factors. Currently, the most widely used program is the Educational Value-Added Assessment System (EVAAS; The SAS Corporation, n. d.-). Some form of the EVAAS has been implemented (or is being considered for implementation) in over 300 school districts in 21 states. The statistical models that underlie VAM approaches such as the EVAAS are complex and incorporate techniques that, in theory, adjust for such factors as preexisting differences in the demographic and academic characteristics of students and the influence of previous schooling on test score growth (Ballou, Sanders, & Wright, 2004; McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004; Sanders, Saxton, & Horn, 1997).

It is very unclear whether a VAM can be used to estimate quantities that can be reasonably interpreted as causal effects (Rubin ~~et al.~~, Stuart, & Zannato, 2004; Briggs & Wiley, 2008). A necessary condition for the use of a VAM to estimate teacher *~~"~~effects~~"~~* is the availability of longitudinal data on a collection of teachers with student test scores that have been linked over time. A statistical model can then be used to estimate the average score increment each teacher has contributed to the achievement of ~~its~~ his or her students in a current year over and above the achievement that had been observed for students in prior years. These *~~"~~increments~~"~~* are not interpretable as causal effects in and of themselves. For this we must establish—for each teacher—a control group of students to represent the average test score increment that would have been observed had students not attended a class with the teacher being viewed as the educational treatment. In the EVAAS, this outcome is represented by the full sample of students across the collection of teachers being analyzed. As a result, value-added ~~"~~effects~~"~~ are estimated and interpreted relative to the average score gain contributed by all schools under analysis. The data employed for a value-added analysis are essentially an extreme version of an observational study in which students self-select the teacher (and by extension, schools) to which they are exposed. A key question of interest is whether different value-added models are better able to adjust for these sorts of selection biases than others.

Formatted: Font: Italic

Formatted: Font: Italic

The results from such a sensitivity analysis were presented by Lockwood et al. (2007) ~~and colleagues in a 2007 study published in the *Journal of Educational Measurement*~~. The authors examined ~~four~~ 4 years of longitudinal data for a cohort of 3,387 students in grades 5 through~~to~~ 8 attending public schools in the state of Pennsylvania from 1999~~-~~ to 2002. Of interest was the sensitivity of teacher effect estimates to the complexity of the VAM being specified. The authors chose four different VAMs in order of the complexity of their modeling assumptions: gain score, covariate adjustment, complete persistence, variable persistence. They also chose five different sets of control variables to include in the VAMs: none, demographics, base year test score, demographics plus base year test score, and teacher-level variables. Finally, they considered one novel ~~variable~~ factor seldom explored in prior VAM sensitivity analyses: the outcome measure. Students in the available sample had been tested with the Stanford 9 assessment across grades 5 ~~to~~ through 8. Upon examining the items contained in the

Stanford 9, Lockwood et al. disaggregated the test into two different subscores as a function of items that emphasized problem solving (40% of the test)~~,~~ and items that emphasized procedures (60% of the test). Having established three ~~variables~~ factors for their sensitivity analysis (type of VAM, choice of covariates, choice of test outcome), the authors estimated teacher effects for each three-~~-~~way ~~variable~~ factor combination and asked the question: W~~w~~hich ~~variable~~ factor has the greatest impact on inferences about a given teacher's effect on student achievement?

What they found was that, by far, the choice of test outcome had the biggest impact on teacher effect estimates. Regardless of the choice of VAM or covariates, estimates of teacher effects tended to be strongly correlated (0.~~-~~8 or higher). On the other hand, the correlations of teacher effects estimates by outcome were never greater than 0.4, regardless of the underling VAM or choice of covariates.

9.1.4 Can Readily Available Standardized Tests Support Causal Conclusions?

Formatted: Heading 3

Formatted: Heading 3, Indent: First line: 0"

I chose the two examples above because they ~~are illustrative of~~illustrate the kinds of evaluative studies that are now being conducted thanks to the testing infrastructure spurred by NCLB. Administrators, parents, and policymakers are naturally going to want to use existing tests to address causal questions about the effectiveness of educational interventions. At this point, I think ~~it is something of an open~~the question of whether the tests are up to the task—regardless of the quality of the underlying study design—is rather open. Imagine that each of the studies described above involved a randomized controlled experiment—the gold standard for estimating unbiased causal effects. This change would mean that in the Ridgway et al.~~-~~ (2000) study, schools were randomly assigned to the CMP or non-CMP curriculum, while in the Lockwood et al. (2007) study, students were randomly assigned both to ~~both~~ schools and teachers. Assume further than the effects estimated in each study were unbiased estimates. Now if each study ~~was~~ were conducted only using the test scores readily available to researchers through state testing programs—ITBS and Stanford 9 math test scores—we would miss a good chunk of the story about the effectiveness of the CMP curriculum and Pennsylvania teachers.

Most schools are eager to implement educational interventions that have been proven to ~~"~~work.~~"~~ To facilitate such decisions, the U.S. Department of Education has established the What Works Clearinghouse (WWC) as source where decision makers can turn to for evidence about a~~n~~ prospective intervention's effectiveness. The WWC is responsible for reviewing the quality of existing studies conducted to evaluate the effects of a wide range of educational interventions. However, such reviews focus almost exclusively on the internal validity of estimated causal effects (Briggs, 2008). Evidence that tests are valid for the causal inferences they are being used to support has been essentially delegated to state departments of education and their test contractors.

Formatted: Body Text, Left, Indent: First line: 0"

9.2 Building a Case for Test Validity: Theory and Practice

9.2.1 *Test Validation in Theory*

Perhaps the most famous and widely cited definition of what is meant by test validity comes from Messick's chapter on validity in the ~~3~~third ~~rd E~~edition of the book *Educational Measurement* (Messick, 1989). Messick wrote, "Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores." (p. ~~XXX~~13). ~~[Derek: Please include page number for this quote.]~~ Messick's contributions to validity theory, building upon the work of Cronbach (1971) and Cronbach and ~~&~~ Meehl (1955), were both influential and somewhat controversial, because he rejected the formerly held trinitarian view of different types of validity (i.e., content, criterion, and construct), and emphasized the view that it is test scores, not the test itself, that are validated. In the process, he redefined the term *construct validity* as a single unitarian concept that encompassed content and criterion-related validity, and made the consequences of testing a fundamental aspect of what is required to establish construct validity.

In the latest edition of *Educational Measurement*, Kane advances what he has ~~called~~ described as "an argument-based approach to validity" (Kane, 1992; 2006, ~~p. XXX~~). ~~[Derek: Please add page number for the quote.]~~ Kane's thesis, consistent in spirit with the perspectives of Shepard (1993), Messick (1989), and Cronbach (1971), ~~and Shepard (1993) [Derek: Please include this source in the references.]~~ before him, is that test validity is a matter of degree, and depends upon the clarity, coherence, and plausibility of any interpretive argument that links test scores to the decisions and inferences for which they are to be used. The essence of the argument-based approach to validation is ~~very~~ appealing: Be clear about how you plan to interpret and use test scores, build a case for why the test in question meets these needs, and defend yourself against alternative cases for why the test is inadequate. On the other hand, as a theory, the approach is incredibly broad and intentionally non-proscriptive.

9.2.2 *Test Validation in Practice*

This view of establishing test validity as the process of integrating different sources of evidence into a comprehensive argument has been formalized in c~~C~~hapter 1 of the *Standards for Educational and Psychological Testing* (~~AERA~~American Educational Research Association~~/APA~~, American Psychological Association, & ~~/NCME~~National Council on Measurement in Education, -1999; ~~[Derek: Please include this source in the references.]~~ hereafter referred to as "*Test Standards for Validity*"). The *Test Standards for Validity* provide five categories of evidence from which an argument for or against the validity of any specific test score inference or consequence could be advanced: (a~~1~~) test content, (b~~2~~) the response processes of test-takers, (c~~3~~) the internal structure of the test,

Formatted: Heading 1
Formatted: Body Text, Indent: First line: 0"
Formatted: Not Highlight
Formatted: Highlight
Formatted: Highlight
Formatted: Font: Italic
Formatted: Not Highlight
Formatted: Not Highlight
Formatted: Not Highlight
Formatted: Not Highlight
Formatted: Not Highlight
Formatted: Not Highlight
Formatted: Not Highlight
Formatted: Body Text
Formatted: Heading 3, Don't keep with next, Don't keep lines together
Formatted: Body Text, Indent: First line: 0"
Formatted: Not Highlight
Formatted: Highlight
Formatted: Highlight

(d4) the relationship of test scores to other variables, and (e5) the consequences of test use. If the *Test Standards for Validity* isare to be taken seriously as a reflection of the consensus position on validity theory, then a critical question is to what extent they its informs the practices of states, especially since NCLB was enacted. Two recent reviews have examined the gap between theory and state practices. Linn (2006) examined the validity evidence used to support test score inferences in the assessment programs of six states: California, Colorado, Florida, Ohio, South Carolina, and Washington. Using information submitted to the U.S. Department of Education as part of the NCLB peer review process (U.S. Department of Education, 2004), Linn compared the validity practices of each state against five categories of validity evidence described in the *Test Standards for Validity*. Linn found that while the states generally provided a great deal of evidence about the content and internal structure of their standardized tests, and about the relationship of scores on these tests with other variables, there was little evidence existed to show that the states were actively investigating the response processes of test-takers and consequences of test use (Linn, 2006). Ferrara (2006) conducted a similar review and concluded that "they types of evidence provided fall far short of current thinking and recent methodological developments relevant to developing validity evidence. Technical reports tend to describe evidence without integrating it into statements about the validity of various interpretations and uses" (Ferrara, 2006, p. 616).

My own analysis of the information and evidence that states make publicly available to support their testing programs have produced results that are consistent with the findings described Linn and Ferrara. However, the fact that there is a gap exists between validation theory and practice does not necessarily imply that tests are being invalidly used for high--stakes purposes. What can be safely concluded is that large-scale standardized tests administered from state to state

- have items that were approved by committees of subject matter experts as being representative of a state's content standards,
- have scores that are suggestive of high reliability, and
- are developed so to avoid obvious cultural biases.

Such information is valuable to be sure. However, these (and other) readily available pieces of information are only links from a what should be a larger argumentative chain of reasoning. One important link that is missing is evidence showingas to the extent to which test scores are sensitive to formal instruction. Such an assumption seems implicit in both the studies by Ridgway et al. (2000) and Lockwood et al. (2007), and would seem to be a central assumption behindto virtually all state tests used to support systems of educational accountability. Yet this does not seem to be an assumption that is beingdoes not seem to be regularly validated.

9.2.3 Test Validation as Causal Inference

It seems to me that one principal reason there is such a gap exists between validation theory and practice — and the reason it is so hard to validate the use of tests for high-stakes inferences — is because the approach outlined in the *Test Standards for Validity* essentially requires us to build an inferential argument by observing effects and

Formatted: Body Text

Formatted: Bulleted text, No bullets or numbering

Formatted: Bullets and Numbering

Formatted: Body Text

Formatted: Heading 3, Don't keep with next, Don't keep lines together

then attributing them to a cause (which is daunting) rather than estimating the effects from a hypothesized cause (which is doable).

Figure 9.3 illustrates a typical psychometric conceptualization of the relationship between test items and a single latent construct underlying these test items. This conceptualization has an implied causal inference, where the idea seems to be that having more or less of the latent construct causes a test-taker to answer a given item correctly or incorrectly. This idea is formalized in item response theory with the conditional expectation $P(X = x_i \mid \theta)$. From this perspective, a necessary condition for establishing test validity is to establish that θ has a causal effect on item responses. The impediment, of course, is that θ is unobserved (and hence not manipulable). As a result, we can only observe differences in the item responses among test-takers~~,~~ and use these to make a causal attribution about θ. So θ is operationally defined only by patterns of item responses. This ~~is~~ result explains why the validity evidence typically provided by psychometricians in the technical reports of state testing programs rely so heavily upon evaluations of test item characteristics: their quality, their intercorrelations, ~~etc~~and so on. A problem with such approaches is that it becomes possible to do analysis that is largely divorced from design. Because no hypotheses are being advanced for what we should expect to observe, almost any finding can be rationalized as acceptable within some bounds for acceptable (and perhaps arbitrary) ranges of item difficulty, point biserials, and reliability.

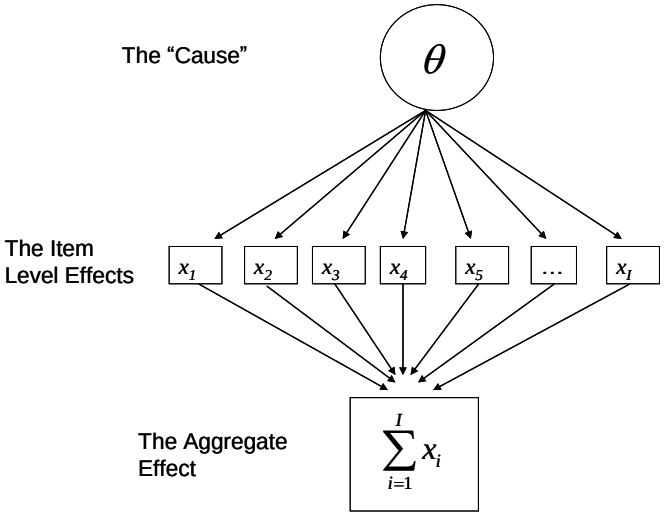


Figure 9.3. Test ~~v~~Validation as ~~c~~Causal ~~i~~Inference.

The notion that causal inference is implicit in test design and validation is not new. This idea can be found in recent manuscripts in ~~the~~ psychometrics literature (c.~~-~~f., Borsboom, Mellenbergh, & Heerden, 2004; Wilson, 2005)~~,~~ and for decades in books and articles on structural equation modeling. However, in my view it is neither feasible nor necessary to model all the causes, latent or otherwise, that influence the item responses of

Formatted: Body Text, Indent: First line: 0"

Formatted: Figure Caption

Formatted: Body Text, Indent: First line: 0"

test-takers on large-scale assessments. In his chapter "Test Validation" from the ~~2nd~~ second edition of *Educational Measurement*, Cronbach (1971) pointed toward a more direct approach when he wrote:

> Experimental interventions in which something is deliberately done to change student scores, as a means of identifying influences to which test performance is sensitive, have been mentioned several times. The treatment may be a change in time limit, a special instruction, etc. The investigator, knowing of what his treatment consists, can predict its effect on the tests; the results confirm or challenge some part of his interpretation of the measuring instrument. (~~Cronbach, 1971, p~~p. 474)

Cronbach was essentially proposing the substitution of an observed and well-understood educational intervention, *Z*, for the hypothesized latent construct θ in Figure 9.3. By "well-understood," I mean that *Z* should have been designed such that not only would exposure to it be expected to have an effect on overall test performance, but also that this effect could be properly hypothesized for specific items or item subsets. That is, if test-developers really understand what is being measured, it should be possible to imagine interventions that would (or at least should) increase the probability of students answering some test items correctly, but *not* increase the probability of answering other items correctly. I illustrate this notion in Figure 9.4.

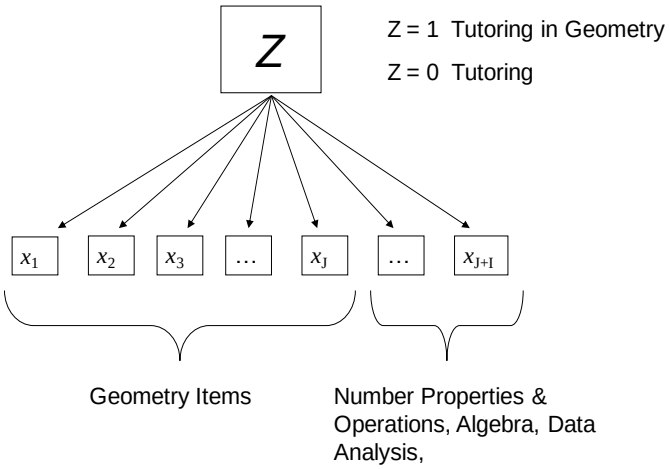


Figure 9.4. Tutoring program as an indirect manipulation of the construct of measurement.

Here we imagine a scenario in which the middle-school students in a state are tested annually on a large-scale assessments of math. The items on the test have been designed to measure different "*content strands*" according to the state's published standards framework, and these strands distinguish between the mastery of number properties and operations, algebra, data analysis, and geometry. Now, if we were to take a sample of students and randomly assign them to either a tutoring program that focused on instruction and practice in understanding geometric concepts (Group 1, *Z*=1), or a tutoring program that focused on algebra (Group 2, *Z*=0), we should expect

Formatted: Block quote para 1, Indent: Left: 0"

Formatted: Body Text, Indent: First line: 0"

Formatted: Figure Caption

Formatted: Body Text, Indent: First line: 0"

Formatted: Font: Italic

that when the test performance of the two groups is compared ~~statistically~~, G~~g~~roup 1 students will have a significantly higher probability of answering geometry items correctly relative to algebra items, and vice-versa for G~~g~~roup 2 students. If we find this to be so, it would seem to bolster an argument that a manipulation of the underlying construct has had an effect on item response probabilities. A competing explanation that would need to be ruled out is that ~~what~~ at least some portion of what the test measures is trivial (~~"~~construct irrelevant) and can be manipulated through savvy coaching techniques (which results in what Koretz & Hamilton (2006) have called ~~"~~*score inflation*~~"~~). If ~~there are~~ no significant differences in the average response probabilities exist between the groups, it would seem to suggest that whatever the test is measuring is not readily manipulable. Again, a competing argument would need to be ruled out: P~~p~~erhaps the tutoring that was implemented differs from what was intended.

Note that in this brief example the central component of a validity argument becomes a matter of estimating effects rather than attributing cause. Of course, much hinges upon the defensibility of substituting ~~"~~Z~~"~~ in place of θ. But in my view, being forced to make and defend this argument ~~this~~ focuses important attention on the intended alignment between what is being taught and what is being assessed. If th~~e~~is substitution of *Z* in place of θ can be defended, then much of the theory and practice of causal effect estimation can be implemented at the item level. The resulting patterns would provide evidence for what a test is~~,~~ and is not measuring. And making item-level inferences ~~This~~ would be possible (though challenging) even when students have not been randomly assigned into tutoring conditions. ~~[Derek: Please replace "this" with a noun, i.e.: this result, this finding…]~~

9.3 Evaluating a Test's Instructional Sensitivity in Practice

~~———~~The provision of supplemental educational services (which I hereafter refer to as tutoring) to low-income students in schools failing to make AYP under NCLB is just now beginning to attract the attention of educational researchers. In my view, it should really be attracting the attention of psychometricians. The tutoring that students are receiving is likely to be the purest form imaginable of teaching to the test. The theory of action behind NCLB and all systems of educational accountability is that a student who has a poor understanding of, say, algebra would have a better understanding if ~~they~~ he or she had instead been exposed to some intervention (i.e., better teaching, more motivation, better diet, etc.~~-~~). It follows from this that for educational accountability to achieve the consequences that are envisioned, there are two necessary conditions: the presence of good interventions~~,~~ and standardized tests that are instructionally sensitive.

~~———~~In Colorado during the 2006-07 school year, approximately 1,500 students in grades 4 through 8 received tutoring beyond their normal school instruction in the subjects of math and reading. All of these students were receiving free lunch assistance, and 94% were Black or Hispanic students. There were many more students in the state in the same grades with the same demographic background and prior test performance who were similarly eligible to receive tutoring~~,~~ but either chose or were unable to take advantage of the tutoring services. Because both groups of students had taken the

Formatted: Font: Italic

Formatted: Font: Italic

Formatted: Highlight

Field Code Changed

Formatted: Highlight

Formatted: Highlight

Formatted: Highlight

Formatted: Heading 1, Left, Indent: First line: 0", Don't keep with next, Don't keep lines together

Formatted: Heading 1, Don't keep with next, Don't keep lines together

Formatted: Body Text

Colorado Student Assessment Program (CSAP) tests in 2006~~,~~ and ~~took them~~ again in 2007, it is possible to estimate an effect of the tutoring. In an evaluation conducted by the OMNI Institute (2008), the tutoring appeared to have no aggregate effect on reading performance~~,~~ and a small effect on math performance. The effect found for math performance was not large enough to move any of the students from a performance level classification of unsatisfactory to proficient. These results are consistent with the few other evaluations of NCLB-mandated tutoring that have been conducted to date (~~Burch,~~ Burch, Steinberg & Donovan, 2007; Vergari, 2007). ~~[Derek: This source is listed differently in the references.]~~ However, while the natural conclusion from such studies is that tutoring programs are largely ineffective, another conclusion must be entertained: P~~p~~erhaps the programs are doing exactly what we would expect, and it is simply the case that the tests are not instructionally sensitive.

How could the principles described in the previous section be applied to this empirical context? To make this example as concrete as possible, imagine we have access to the full population of grade 5 students in a single Colorado school district during the 2008-09 school year. A subset of these students ~~were~~ was eligible to receive tutoring services because they were low income and their schools failed to make AYP. To keep things simple in this illustration, we will focus just on math outcomes. ~~There are~~ ~~r~~Roughly 100 items are administered on the grade 5 CSAP math exam, and these have been mapped evenly into five designated content standards according to the state's department of education (number sense, algebra, statistics, geometry, and problem-solving). A first order of business would be to determine, through inspection of curricula or other analysis, the alignment between the tutoring programs and the CSAP math test. Does the program spend equal amounts of time on instruction that would map to each of the five item sets found on the CSAP? (If the tutoring company is being strategic, one might expect them to devote greater energy to the content with difficulty closest to the performance threshold that demarcates "~~"~~proficiency~~"~~.) From this analysis a program-specific hypothesis can be generated about the types of items that should be most sensitive to tutoring. Now assume we have at least two tutoring programs to compare that have been determined to differ significantly in their relative alignment with the CSAP test[1]. In P~~p~~rogram 1, a student has been exposed to a program with the greatest relative alignment to the 40 items emphasizing an understanding of number sense and algebra. In P~~p~~rogram 2, a student~~s~~ has been exposed to a program with the greatest relative alignment to the 40 items emphasizing an understanding of statistics and geometry. Given such information, w~~W~~e can proceed to empirically compare the probability of correct item responses as a function of tutoring exposure after conditioning on math performance in prior grade(s).

One straightforward way this could be done would be to use the Mantel-Haenszel procedure described by Holland & Thayer (1988) for use in the context of diagnosing potential symptoms of item bias. Or, we could use logistic regression ~~techniques~~ and an approximation technique (c.-f., Swaminathan & Rogers, 1990) to estimate the area between curves as a function of tutoring exposure. Conditional on prior ability, students receiving more tutoring in number sense and algebra should outperform their

[1] It would also be possible to compare a single tutoring program to a control condition of no tutoring, but this comparison would introduce a clear source of bias in the sense that students enrolled in tutoring are likely to be more motivated than those who are not.

Formatted: Highlight

Formatted: Highlight

Formatted: Body Text, Indent: First line: 0"

counterparts receiving more tutoring in statistics and geometry on these test items, and vice-versa. Provisional conclusions about the instructional sensitivity of the test would hinge upon the results from these analyses. If the test appears to be instructionally sensitive, it bolsters the validity of its high-stakes use within an accountability system.

To be sure, many details of this approach would need to be ironed out:~~.~~

- How big must an item-level difference between groups be before it is considered practically significant?
- Should the results be aggregated (for example, summed across all number sense and~~&~~ algebra items) or evaluated item by item for salient trends?
- Should an estimate of the current test score be used as a conditioning variable or only prior test scores? Should all available test score information be included? (Note: this increase in dimensionality could be reduced through propensity score estimation.~~-~~)
- When students have not been randomly assigned to tutoring groups, what other variables are available for inclusion in the conditioning set?

Many of these questions have already been raised (and addressed) in the psychometric research literature on differential item functioning ~~techniques~~ (DIF) techniques. An evaluation of DIF is standard practice for testing companies, but its interpretation is often highly equivocal because the categorical grouping variables employed are usually demographic. In contrast, ~~for the present test validation context~~ the results are more readily interpretable for the present test validation context because the grouping variable is a manipulable treatment that serves as a proxy for the construct of measurement. While it is true that differences in average response probabilities might be due to selection bias (depending upon the reasons that some students choose to enroll in tutoring programs), a mitigating factor is the availability of longitudinal data and the fact that the students eligible for tutoring are, by definition, from low~~-~~-income households. ~~I~~Furthermore, when the item-level performance of students in different tutoring programs is being compared, one might also be willing to assume that, on average, both sets of students are similarly motivated relative to students ~~that~~ who were eligible for tutoring but did not enroll.

9.4 Some Final Comments

An important impetus for the test validation design proposed above is that ~~there needs to be~~ a closer connection needs to be developed between the ways tests are designed and scores are interpreted. By looking for what are essentially causal effect estimates at the item level, we commit ourselves to an understanding of what we think is being taught in schools~~,~~ and what specific item sets we think will capture this learning. States such as Colorado should be able to say, for example, "T~~t~~he principle obstacle to being classified as proficient in mathematics as of grade 5 is an understanding of basic concepts in geometry and their application to solve measurement problems. So this should be the focus of our tutoring programs.~~-~~" If tutoring programs were to then respond by teaching geometric concepts and applications, we should expect to see causal effects

Formatted: Bulleted text, No bullets or numbering

Formatted: Bullets and Numbering

Formatted: Body Text, Indent: First line: 0"

Formatted: Heading 1, Left, Indent: First line: 0"

Formatted: Heading 1, Indent: First line: 0"

on the associated geometry items, but not on items that focus, for example, on number sense. If we do, this is strong evidence in favor of test validity. If we do not, then I think we need to carefully consider that beyond the possible explanation that the tutoring is ineffective, there is a possibility that the existing test is not valid for the high-stakes inferences inherent in accountability systems.

In conclusion, I think we can gain much more traction in validating the use of test scores for high-stakes inferences if we make our causal hypotheses complex, but keep our analyses relatively simple. The evaluation of tutoring programs under NCLB provides a unique opportunity for implementing this idea. In my view, these are the kinds of validation studies that it would be easy to convince states to do because they are at once theory -driven and pragmatic. T theory-driven because you have to know what it is your tutoring purports to teach and your tests purport to measure, but pragmatic because it they may save states millions of dollars being spent on tutoring that does not help, or on tests that are invalid for their proposed uses. When tests must be validated for use in supporting high-stakes causal inferences, the traditional sources of validity evidence are necessary, but not sufficient. If we wish to avoid causal inferences that are careless, we proceed with business as usual at our own peril.

References

Formatted: Heading 1, Left, Indent: First line: 0"

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). Validity. In *Standards for educational and psychological testing*. (pp. 9-24) Washington, DC: American Educational Research Association.

~~American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). [Derek: Please include this source in the references.]~~

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics, 29*(1), 37–65.

Formatted: Reference, Indent: Left: 0", First line: 0"

Briggs, D. C. (2008). Synthesizing causal inferences. *Educational Researcher, 37*(1), 15–22.

Formatted: Font: Italic

Briggs, D. C., & Wiley, E. (2008). Causes and effects. In L. Shepard & K. Ryan (Eds.), *The Future of Test-based educational accountability*, ~~L. Shepard & K. Ryan (eds)~~. New York, NY: Routledge.

Borsboom, D., Mellenbergh, G., & Heerden, J. (2004). The concept of validity. *Psychological Review, 111*(4), 1061–1071.

Burch, P., Steinberg, M., & Donovan, J. (2007). Supplemental educational services and NCLB: Ppolicy assumptions, market practices, emerging issues. *Educational Evaluation and Policy Analysis, 29*(2), 115–133. [Derek: This source is cited differently in the text.]

Cronbach, L. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). Washington, DC: American Council on Education.

Cronbach, L., & Meehl, P. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281–302.

~~Cronbach, L. (1971) Test validation. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed., pp. 443-507). Washington, DC: American Council on Education.~~

Ferarra, S. (2006). Standardized assessment of individual achievement in K-12. In R. L. Brennan (Ed.-), *Educational mMeasurement* (4th ed.- ~~). (~~pp. 579–621). Westport, CT: American Council on Education/Praeger.

Holland, P. W. (1986). Statistics and causal inference~~,~~ (with discussion and rejoinder). *Journal of the American Statistical Association, 81*, 945–970.

Holland, P., & Thayer, D. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. Braun (Eeds.), *Test vValidity* (pp. 129-145)~~XX-XX~~). Hillsdale, NJ: Lawrence Erlbaum Associates~~City, ST: Publisher.~~ [Derek: Please add page numbers, city, state, and publisher.].

Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*(3), 527–535.

Formatted: Font: Italic

Formatted: Not Highlight

Formatted: Not Highlight

Formatted: Not Highlight

Formatted: Not Highlight

Formatted: Highlight

Formatted: Reference, Indent: Left: 0", First line: 0"

Formatted: Font: Italic

Formatted: Font: Italic

Formatted: Not Highlight

Formatted: Not Highlight

Formatted: Highlight

Formatted: Highlight

Formatted: Font: Italic

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.). (pp. 17–64). Westport, CT: American Council on Education/Praeger.

Koretz, D., & Hamilton, L. (2006). Testing for accountability in K–12. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.). (pp. 531–578). Westport, CT: American Council on Education/Praeger.

Linn, R. (2006). *Validity and reliability of student assessment results.* Unpublished manuscript. Paper prepared for the Urban Institute in connection with a contract with the U. S. Department of Education for work related to No Child Left Behind assessment evaluation. Preliminary dDraft.

Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V., & Martinez, J. F. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement*, *44*(1), 47–68.

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics*, *29*(1), 67–101.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed.). (pp. 13–103). New York, NY: American Council on Education and MacMillan Publishing Company.

Messick, Cronbach, & Shepard. (1993). [Derek: Please include this source in the references.]

No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 Stat. 1425 (2002). Retrieved July 22, 2007, from http://www. ed. gov/legislation/ESEA02/

OMNI Institute. (2008). *Evaluation of Supplemental Educational Services: 2006-07 Academic Year Data.* Unpublished rReportmanuscript.

Formatted: Font: Italic

Formatted: Font: Not Italic

Formatted: Reference, Indent: Left: 0", First line: 0"

Formatted: Font: Italic

Formatted: Reference, Indent: Left: 0", First line: 0"

Formatted: Font color: Black, Not Highlight

Formatted: Font color: Black

Formatted: Font: Italic

Formatted: Font: Italic

Formatted: Font: Italic

Formatted: Font: Italic

Formatted: Font: Italic

Formatted: Font: Italic

Formatted: Font: Italic

Ridgway, J., & Schoenfeld, A. (1994, month). *Balanced aAssessment: Designing assessment schemes to promote desirable change in mathematics education.* Keynote paper for presented at the EARLI Email Conference on Assessment. City, ST. [Derek: Please add month, city, and state of conference location.]

Ridgway, J., Zawojewski, J., & Hoover, M. (2000). Problematising evidence-based policy and practice. *Evaluation and Research in Education*, *Vol. 14*(3, &4), 181–192.

Rubin, D., Stuart, A., & Zannato, E. (2004). A potential outcome view of value-added assessment in education. *Journal of Educational and Behavioral Statistics*, *29*(1), 103–116.

Sanders, W. L., Saxton, A. M., & Horn, S. P. (1997). The Tennessee value-added assessment system, a quantitative, outcomes-based approach to educational measurement. In Jason Millman (Ed.), *Grading teachers, grading schools, Is student achievement a valid evaluation measure?* (pp. 137–162). Thousand Oaks, CA: Corwin Press.

Shepard, L. (1993). Evaluating test validity. *Review of Educational Research*, Vol. 19, 405-450. Washington, DC: American Educational Research Association. XX–XX. [Derek: Please add page numbers for this chapter.]

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*(4), 361–370.

The SAS Corporation. (n.d.). *Schooling eEffectiveness, SAS® EVAAS® for K–12*. Retrieved January 4, 2006, from http://www.sas.com/govedu/edu/services/effectiveness.html

Formatted: Highlight
Formatted: Font: Italic
Formatted: Highlight
Formatted: Highlight
Formatted: Highlight
Comment [DB2]: I can't find this exact information--this is a secondary citation drawn from the article by Ridgway et al (2000) [see below]
Formatted: Font: Italic
Formatted: Reference, Indent: First line: 0"
Formatted: Reference, Indent: Left: 0", First line: 0"
Formatted: Highlight
Formatted: Highlight
Formatted: Font: Italic
Formatted: Reference, Indent: First line: 0"

U.-S. Department of Education. (2004). ~~No Child Left Behind.~~ *Standards and assessments peer review guidance: Information and examples for meeting requirements of the No Child Left Behind Act of 2001*. Washington, DC: U.-S. Department of Education, Office of Elementary and Secondary Education~~, April 28~~.

Vergari, S. (2007). Federalism and market-based education policy: T~~t~~he supplemental educational services mandate. *American Journal of Education*, 113, 311–3-39.

Wilson, M. (2005). *Constructing measures: A~~a~~n item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum Associates.

Formatted: Reference, Indent: Left: 0", First line: 0"