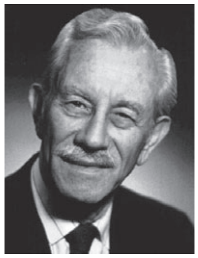
Measuring the Power of Learning.®

A TALE OF TWO MODELS: SOURCES OF CONFUSION IN ACHIEVEMENT TESTING

By Mark D. Reckase

WILLIAM H. ANGOFF MEMORIAL LECTURE Series

William H. Angoff
(1919 – 1993)



William H. Angoff (1919 – 1993) was a distinguished research scientist at ETS for more than 40 years. During that time, he made many major contributions to educational measurement and authored some of the classic publications on psychometrics, including the definitive text, "Scales, Norms, and Equivalent Scores," which appeared in Robert L. Thorndike's Educational Measurement. *Dr. Angoff was noted not only for his commitment to the highest technical standards but also for his rare ability to make complex issues widely accessible. The Memorial Lecture Series established in his name in 1994 honors Dr. Angoff's legacy by encouraging and supporting the discussion of public interest issues related to educational measurement. These lectures are jointly sponsored by ETS and an endowment fund that was established in Dr. Angoff's memory. The William H. Angoff Lecture Series reports are published by the Center for Research on Human Capital and Education, ETS Research and Development*.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

A Tale of Two Models: Sources of Confusion in Achievement Testing

Mark D. Reckase
Michigan State University, East Lansing, MI

A common interpretation of achievement test results is that they provide measures of achievement that are much like other measures we commonly use for height, weight, or the cost of goods. In a limited sense, such interpretations are correct, but some nuances of these interpretations have important implications for the use of achievement test results. This paper will contrast two different theoretical underpinnings for the interpretation of educational assessments results as measurements. One of these theoretical views comes from initial attempts in psychology to measure the amount of a trait that was exhibited by a person. The other theoretical view comes from early work in education to measure how much of a desired curriculum was acquired by students. At times, these views conflict with each other and lead test developers and policy makers to ask for the impossible. After summarizing the two theoretical positions, the areas of conflict will be discussed. Finally, some recommendations will be given for what can be done to clarify the issues and minimize the problems that result from using conflicting theoretical frameworks.

Keywords Achievement testing; continuum model; domain sampling model; test design; test development; educational measurement; psychology

doi:10.1002/ets2.12171

Preface

The 16th William H. Angoff Memorial Lecture, *Do Educational Assessments Yield Achievement Measurements?*, was presented by Dr. Mark D. Reckase, University Distinguished Professor Emeritus, Michigan State University, at The National Press Club in Washington, DC, on March 29, 2017. In this paper based on his lecture, Dr. Reckase addresses two perspectives on test design: the psychology perspective and the educational perspective. In the psychology perspective, the goal of instrument development is to produce items or tasks that will estimate a student's location along a specific continuum. In the educational perspective, the goal is to estimate what proportion of a specific domain has been acquired by students. On the surface, the two approaches appear similar: Both select a set of test items to administer; both are designed to estimate a single score. However, Dr. Reckase cautions that mixing these two development concepts can lead to confusing interpretations and weakly supported results.

Dr. Reckase also addresses the changing perception of measurement by pointing out that the visible means by which the previous generation learned to measure time and space in elementary school are no longer apparent in the 21st century. Distance is now measured by GPS instead of a measuring tape; time is marked by a digital readout rather than hands on a clock. Test measurement, he tells us, follows the same pattern: Achievement is not measured by a sum of scores from test items but by various statistical models unseen to the end users of the results.

In his conclusion, Dr. Reckase offers suggestions to balance the approach to test design with the use of results in order to gain a more accurate picture of what today's students know and can do.

The William H. Angoff Memorial Lecture Series was established in 1994 to honor the life and work of Bill Angoff, who died in January 1993. For more than 50 years, Dr. Angoff made major contributions to educational and psychological measurement and was deservedly recognized by the major societies in the field. In line with Dr. Angoff's interests, this

Corresponding author: M. D. Reckase, E-mail: reckase@msu.edu

lecture series is devoted to relatively nontechnical discussions of important public interest issues related to educational measurement.

Ida Lawrence
Senior Vice President
ETS Research & Development

Introduction

The motivation for writing this report is my sometimes frustrating experience of either orchestrating achievement test development or reviewing the achievement test development processes used by others. These experiences are frustrating because approaches to test development sometimes appear to have conflicting goals that make it difficult to produce a quality test that provides the desired information. The purpose of this report is to lay out what I think are at least partial causes for the conflicts and my resulting frustration. These causes are presented with the hope that they will clarify issues and thereby help test developers avoid some of the mistakes I and others have made when approaching the challenge of producing sound tests for assessing educational achievement.

At the most basic level, this report is about attempts to measure student achievement, so it is important to consider what it means to measure something. Measurement is one of those things that on the surface seem very simple, like stating that $1+2=3$, but are very difficult to treat in a rigorous mathematical way. There are very difficult theoretical texts on the topic of whether test results can be treated as measurements, such as the three-volume set by Krantz, Luce, Suppes, and Tversky (1971) with various orders of authorship (Luce, Krantz, Suppes, & Tversky, 1990 and Suppes, Krantz, Luce, & Tversky, 1989). This report is not going to consider measurement in that technical way because it is more focused on how persons who set testing policy or who are consumers of test results think about measurement. Of course, I do not know how they think for sure because I am not inside their heads, but I have observed people in those positions for many years, and I will present some hypotheses about their general understanding of measurement concepts.

Measurement in Everyday Life and the Technology of Measurement

Most of us have learned about measurement through our elementary school education. We have learned that if you want to measure the length of a line or some physical object, a ruler or tape measure is used. And we are taught how to place the ruler or tape measure next to an object and read the numerical value of length from the markings on the ruler or tape. We are also taught how to measure weight (or more accurately, mass) using a spring scale or a balance. And very early on, we are taught about time. When I went to school, we learned to measure time using the change in locations of hands on a clock. These early lessons seem to be very effective, and adults in policy positions or who are consumers of test results are still very adept at measuring tangible things. Making marks on a wall to show the growth in height of a child seems to be a very compelling visual metaphor for making measurements and noting changes in the thing measured.

Some interesting properties are present in these familiar measurements. Length and weight have easy-to-understand zero points. They also have easy representations of equal units through the concatenation of identical pieces. Time is different, however. Although we can count days as time units, it is not clear what a true zero point is for days. Instead, we define arbitrary zero points, such as our day of birth or some other historical event. Time units come from things that we believe are regular, such as orbits around the sun or swings of a pendulum. But, I have learned that even my digital clocks seem to drift from the "official" time for the country. Maybe the regular events are not as regular as is generally believed.

It is an interesting observation and an important point for the ideas presented here that the measurements that we encounter daily for length and weight and time often no longer have a connection to the physical processes we learned about in elementary school. For example, when I go for a daily walk, I use an application on my cell phone to determine the distance that I have walked. Clearly, no ruler or tape measure is involved, but I accept the result that shows up on the screen on my phone as being reasonably accurate. However, I have no idea how it is done. I have some theories that it has to do with GPS satellites and there is probably some geometry involved in getting the distance between successively estimated points, but before writing this report, I did not think very much about it.

Similarly, when I weigh something on my kitchen scale, I get a digital readout, but I do not really know how it works. I suspect that a crystal or a piece of metal changes electrical resistance when pressure is applied and that change in resistance is related to the amount of pressure/weight. Then an electro/physical model connects the weight of the object to the digital presentation of the estimated weight. The case is similar for time. Most of my clocks are now digital, and I think that time is measured by the number of vibrations of a crystal when an electrical current is applied. This guess is based on something I read years ago. However, I am fairly sure that a physical model connects some electrical phenomenon to the passage of time and is used to count units of time.

The point of this discussion is to indicate that even though our understanding of measurement of physical objects and time is based on simple physical measurements, the reality in our current technological society is that these measurements are made with complex models that are not readily apparent to the typical consumer of the results of the measurements. This observation is important because the measurements that come from tests have the same characteristics. The general public learns about tests as measurements in those same elementary school classrooms from the experience of taking and interpreting the tests administered to them by their teachers. They learn that the test score is the sum of the scores from each test item. They may also have learned the time-honored standard of 70% of maximum being the percent of correct responses needed to pass the test. But, like the physical measurement processes, large-scale tests do not obtain scores using the methods employed in elementary school. Instead, various statistical models are used to estimate a student's location on a scale and that estimated location can be transformed to something like the summed score used in elementary classrooms if it is desired. However, usually the results are reported on a different scale. For some time, the methods used by large-scale testing programs were called "model-based measurement" (e.g., Embretson & Reise, 2000, Chapter 3) to emphasize the distinction with procedures based on the sum of item scores.

For this report, achievement measurement will be thought of as somewhat analogous to the measurement of time. Persons are located on a scale relative to a reference point, and we can determine if one person is farther from the reference point than another. I will elaborate on this later in the report.

My Personal Journey to Learn About Achievement Measurement

To explain my own personal view of achievement measurement, I need to give some of my personal history. I do not mean to suggest that my training in the area is any better than others in the field; it is simply different, and that has led me to certain insights.

My training in this area probably began at the elementary and secondary levels as was the case for the general population, but it came to a focus at the university level. I went to the University of Illinois, Champaign/Urbana as an undergraduate student expecting to major in mathematics. But, after a year of taking courses that culminated in one from the head of the department teaching out of a book called *A Course of Pure Mathematics* (Hardy, 1960), I decided that pure mathematics was too abstract for me. So, I tried English literature, human physiology, and philosophy (also too abstract) before settling on psychology with a special interest in experimental social psychology and attitude change. I particularly remember a course taught by Professor Martin Fishbein that included material on creating attitude scales. To investigate attitude change, it was important to know where a person was on an attitude continuum at one point in time and then determine where he or she was after some event or intervention. This was the beginning of my interest in psychological measurement.

After finishing undergraduate work in psychology at the University of Illinois, I went to Syracuse University to pursue a doctorate in psychology with specialization in attitude change. But, my assigned advisor and head of the department, Professor Eric Gardner, told me there were no assistantships available in social psychology, so he was going to put me into psychometrics and statistics for a year, and if I did not like it, I could change later. Well, I did like that program of study and never thought again about returning to social psychology. I learned psychometrics from a psychological perspective and worked on a research project called the Adult Development Study (Monge & Gardner, 1970) where we produced many tests to gage the cognitive abilities of the adult population in the United States that had completed formal schooling. The information we gained was expected to help determine how to retrain people as the job market changed. The result was that I gained a lot of practical experience in test design and development from a psychological perspective.

Because of complications due to time in the U.S. Army after completing my doctorate, I was looking for an academic job in January rather than at the beginning of the academic year. Professor Gardner knew of a position in educational psychology at the University of Missouri—Columbia that had not been filled the previous September. I applied and was granted an interview. That interview presented me with the first information that there might be a difference between test design and development from the educational perspective and the approach that I learned in my psychology program. I still remember the interview question, “What do you think about criterion-referenced testing?” I knew nothing about criterion-referenced testing. I thought they were asking about criterion-related validity and answered the question from that perspective. My answer must have made my level of ignorance very clear, but they hired me anyway. Perhaps they worried that they would lose the position if it was not filled during that academic year.

My first teaching assignment was a large lecture course on classroom assessment for prospective teachers. I knew only a little about this topic, but with the help of another faculty member, Terry TenBrink, who had taught the course for many years and who shared his lecture notes and his book (TenBrink, 1974), I learned the material faster than the students in the class. This course was taught five times a year, and I taught it for 5 years before arranging for my graduate students to teach it on occasion. Over that period, I searched for a textbook for the course that balanced my view of measurement and the requirements of the course. Eventually, I settled on *Measurement and Evaluation in Education and Psychology* by Mehrens and Lehmann (1991).

Teaching a course 25 times in 5 years is one way to gain thorough knowledge of the material. During that time, I also began doing work on computerized adaptive testing (CAT) and used a CAT for retesting students who did not perform well on their paper-and-pencil examination for the testing course. Developing a CAT for this course examination raised concerns about the unidimensionality assumption of item response theory (IRT), and those concerns led to my 1979 paper on how strong the first dimension from a test needed to be to support the use of an IRT-based scaling approach (Reckase, 1979). The conflict between the psychometrics I learned as a graduate student and the characteristics of achievement tests was beginning to become evident.

I spent about 10 years at the University of Missouri and focused research efforts on CAT. Then, in 1981, I applied for a position at American College Testing (now ACT, Inc.) and was hired as director of resident programs. This position was the head of the department that developed the testing programs owned by ACT, including the college entrance examination. The department was also responsible for the ACT Proficiency Examination Program (PEP) (American College Testing Program, 1976), which was a course credit by examination program.

The experience with PEP taught me that persons who run testing programs tend to want more from the test than it was designed to produce. The results from this testing program were generally used to make a pass/fail decision. If a cut score on the test’s score scale was reached or exceeded, examinees would get credit for the corresponding course; if the score on the test was less than the cut score, no credit was awarded. The test worked well for that purpose. But, those who were responsible for testing policy wanted to give those who did not pass some information to help them prepare for a retake. They wanted subscores related to the test content.

So, here was my first instance of giving people what they wanted. We studied the test to see how many items would be needed to get a score that was reliable enough to report and then determined how many meaningful subscores could be reported for each test. We settled on 12 items as the minimum number needed based on the rule of thumb that a minimum reliability of .7 would give meaningful results. Further, we decided to only report three levels for the subscores: less than required, more than required, and not significantly different from the cut score. I do not know if these subscores were helpful for the examinees that did not pass.

For the ACT Assessment (American College Testing Program, 1989), we started using IRT for test design and construction even though the test was equated using the number-correct score. The tests in that program were designed to represent what was common in curricula across the United States for the particular subject matter areas that were considered. The tests were not designed to give a representative sample of items but, rather, a sample of items that would be predictive of performance in entry level courses at a college or university.

Later at ACT, our research team designed a CAT for placement in entry level college courses at 2-year colleges called COMPASS (ACT, Inc., 2006). When designing that testing program, we had to confront the problem of using a unidimensional IRT model as the basis for the CAT, but the test specifications had to cover complex areas of achievement. We addressed the problem by making several unidimensional adaptive tests with software that would branch seamlessly from one to the other. An examinee could start with a test of arithmetic skills and knowledge and, if they performed well, quickly get branched to a test of algebra skills and knowledge. In mathematics, we developed five different tests based on unidimensional IRT models with flexible branching from one to another.

Two Perspectives on Test Design and Development

After many years of not only working on the development of achievement tests and teaching teachers how to produce them, but also with training in psychological testing and completing research on IRT and CAT, I began to realize that there are two different perspectives on test design and development, and these perspectives are not consistent with each other. This discovery is particularly troubling because many books that are used for teaching testing courses have keywords in their titles, like "educational and psychological testing," that do not highlight the differences in the approaches to test design and development. Instead, the titles imply that educational and psychological testing are the same thing or are, at least, very similar. The rest of this report describes the different perspectives on test design and development and then gives some examples of the problems that can arise if users of test results do not have a clear understanding of the differences.

Perspective 1: The Psychological Perspective

As I indicated previously, the initial understanding about test design and development that I gained through my undergraduate and graduate training in psychology was that persons are different on many different characteristics. These may be called traits or abilities, but there is a common idea that for each of the characteristics individuals differ in degree or amount. Each characteristic is thought to have a continuum that is defined from low to high or from most negative to most positive. This conception probably stems from the early work in psychophysics when many measurements were of reaction time and differences in perceptions of tones, strength, and so on. See Thorndike (1904) and Brown (1911) for a discussion of the requirement for measurement using homogeneous tasks and the idea of ordering persons on a continuum.

From this perspective, the goal of instrument development is to produce items or tasks that will give information to support accurate estimation of a person's location along the target continuum (see Wissler, 1901, for descriptions of these tests). Later, scales were expanded to many other areas using ratings on bipolar scales with opposite adjectives at the extremes, such as slow versus fast or cold versus hot (Osgood, Suci, & Tannenbaum, 1957). A useful visual metaphor for the continua behind the scales for psychological tests is a long railroad track winding through the countryside. In a sense, the task of a psychological test is to ask questions to find the location of a person who is on a train on that track. This metaphor will be expanded with an example later in this report.

Perspective 2: The Educational Perspective

The approach to test design and construction taken in educational settings is based on the idea that a domain of knowledge and skills is the target for instruction. The goal is to estimate what proportion of that domain has been acquired by students. This approach to test design and development has a very long history. As early as 1897, Rice (1897) presented descriptions of tests composed of items that were sampled from domains in spelling. Cornman (1902) replicated Rice's work and described the process:

> Each column test consisted of fifty words selected from a "Review List of Difficult Words" for the particular grade to be tested, found in a modern spelling book. The review words (about five hundred in a list) were arranged in alphabetical order, and the fifty words for the first list were selected by taking the first, fifth, ninth, thirteenth word, etc.: those for the second list, by taking the second, sixth, tenth word, and so on. In this way, the lists were secured presenting approximately equal degrees of orthographical difficulty. (pp. 62–63).

For this spelling test, the domain of content was very well defined and systematic sampling was used to get representative samples for multiple test forms. Forty-two years later, Greene, Jorgensen, and Gerberich (1944) described a similar sampling process for developing achievement tests:

> Tests make no pretense of measuring every skill, ability, fact attitude, etc., which the pupils acquire as outcomes of instruction. Such comprehensive measurement would be impossible with present measurement techniques and moreover would be relatively wasteful of time and effort. As a substitute, the same procedures of sampling as are used in many fields have been adapted to test construction. Just as a grain buyer samples a carload of wheat by taking

> samples in order to obtain a measure of quality for the whole carload, a test constructor measures the educational attainments of pupils by constructing test items which represent widely the types of pupil outcomes expected and accepts the scores resulting from their use as representative of the pupils' relative achievements for the entire area sampled by the test items. *Adequacy is the degree to which a test samples sufficiently widely that the resulting scores are representative of relative total performance in the areas measured*. (pp. 63–64).

Current thinking about achievement testing is not very different than it was in 1944. Downing (2006) continued to describe the main task of test development as obtaining a representative set of tasks from the domain; however, he now referred to a domain of skills and knowledge, such as those needed to practice medicine, as a "construct":

> The validity of inferences for achievement test scores rests primarily and solidly on the adequacy and defensibility of the methods used to define the content domain operationally, delineate clearly the construct to be measured, and successfully implement procedures to systematically and adequately sample the content domain. (p. 7)

The use of the term *construct* here adds some confusion between the continuum concept used in psychological testing versus the domain sample used as the basis of achievement testing. On one hand, the sentence implies a single thing, "the construct," but it also emphasizes the sampling to the content domain.

Developing Tests Using the Two Perspectives

To emphasize the difference between the two perspectives for test design and development, idealized test development processes will be presented for each. These examples are simplified for the purposes of this report. Test development is a complex and detailed process that would take pages of text for adequate description. However, the important points about the differences can be made using these idealized examples.

Developing a Test Using the Continuum Model

Developing a test using the continuum model first requires the definition of the continuum. When I worked for the Adult Development Study at Syracuse University, we hypothesized that adults would differ on many different continua and we attempted to develop tests for each of them. An example is a test called *Finance*. It was hypothesized that as individuals gained life experience, they would also gain knowledge about finance, which was considered as a continuum. There was no conception of what a domain for "finance" might be. Instead, the goal was to create test items, in this case multiple-choice items, that would help locate persons along a continuum of financial skills. To do this, items were developed and field-tested to determine which would be highly correlated with the hypothesized continuum. Those that were highly correlated with the total score as a proxy for the continuum were considered and a set that spanned the range of the continuum that was of interest was used to create the short test of the variable, Finance.

A fanciful example that I use with my classes may clarify this point. Suppose that you have a friend who is in Washington DC, the location of my talk on this topic. That friend is someplace on K Street, and you want to meet him. However, your friend damaged his cell phone, and it can only receive text messages and respond with a Y for "yes" and N for "no." What questions would you text to your friend to find out where he is located on K Street? Figure 1 shows a map of Washington DC with arrows pointing to locations on K Street. The analysis we do in my classes shows that if a good set of questions is asked, such as "Are you west of 12th Street?," and your friend always gives an accurate answer, then you can locate that person with high accuracy after about 11 questions. That is, you would be close enough to your friend to see him. However, if your friend does not know the answer to your question—there is no option like "I do not know"—he must give a random response. This scenario is like guessing on a test. This random answer adds noise to the estimation with the result that more questions will be needed to get the same level of accuracy.

From this example, it should be clear that you would ask a different set of questions if your friend was near Rock Creek Park (the west end of K Street) than you would if he was near Union Station (the east end of K Street). It would not make sense to force questions about the area around Union Station if there was already sufficient information to know that your

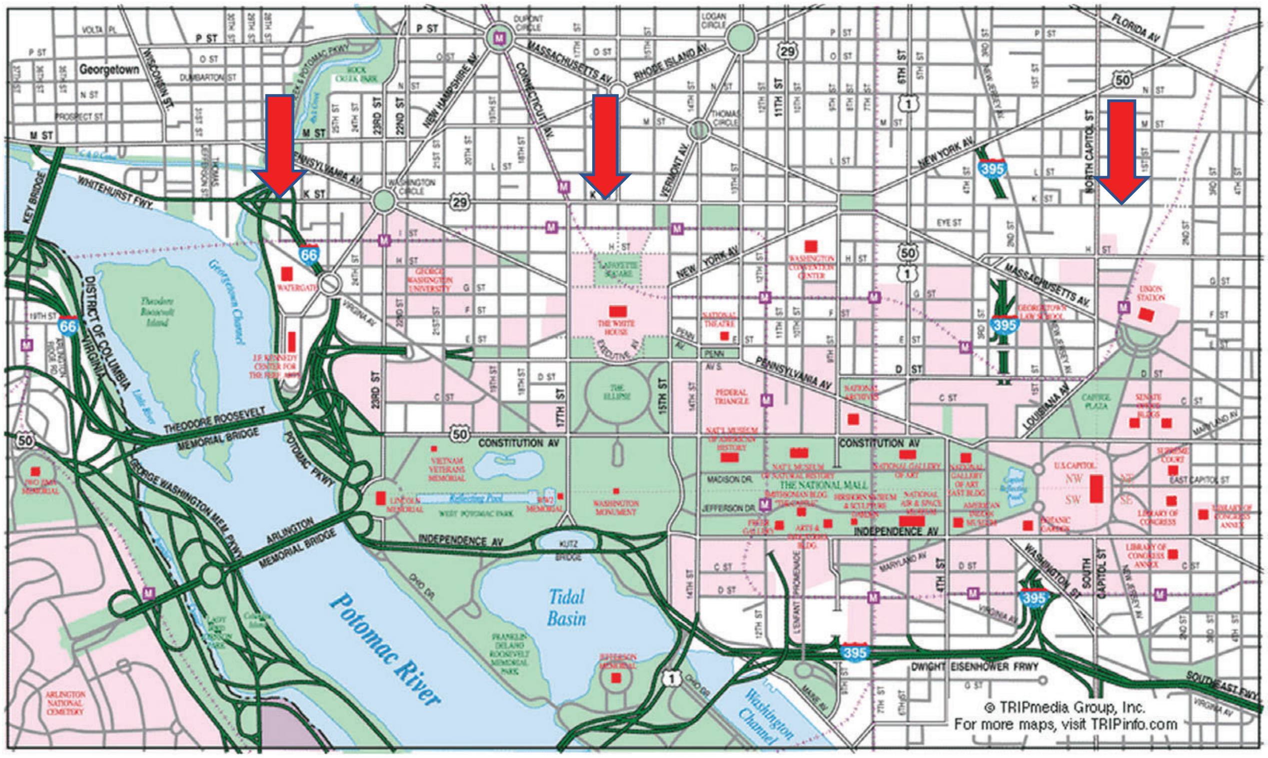


Figure 1 Three possible locations of a person on K Street.

friend was near Georgetown (west end). This type of questioning is the basis for CAT, and it shows a strong distinction between building a test to find a location along a continuum versus a domain sample. You certainly would not want to define a domain of all places along K Street and get a representative sample of questions and then ask all of them. This approach would be inefficient and tedious. Even if the questions were being developed for several different persons, it would be best to think about questions evenly spaced along the continuum rather than randomly sampling them from the domain of all possible questions that could be asked.

Developing a Test Using the Domain Sampling Model

The domain sampling model starts with a detailed description of the domain. This description should be precise enough so that users can tell what is in the domain and what is not. The second step is to develop a test design that will yield a representative sample from the domain (often called a table of specifications). The table specifies a stratified sampling plan using subcontent domains as strata. The goal is to sample from the domain to get an estimate of the proportion of the entire domain that a person has acquired.

In theory, items could be developed to cover the entire domain, and the test could be developed by randomly sampling from the full set. However, that method is seldom used because it is too costly. Instead, items are developed for each stratification cell, and it is assumed that those in the cell represent the full domain of items that could be in the cell. Lately, a substantial effort has been made through what are called alignment studies to check if the set of items represents the domain (see Porter, 2002, for example).

A fanciful example of the domain sampling approach to test development was given by Bock, Thissen, and Zimowski (1997). They randomly selected 100 words from a secretary's spelling guide (Leslie & Funk, 1935) that contained 25,000 words. The percent correct was considered as an estimate of the percent of the total of 25,000 items that a person could spell correctly. This example is fanciful because actual practice with a domain sampling approach uses a more complex stratified sampling approach. For the spelling example, the total sample of 25,000 words might be stratified by part of speech and length of word. Then, the random sample might be drawn from individual stratification cells with the number drawn from each cell based on the proportion of the 25,000 words that were in each cell.

When using this type of test development process, item analyses are still performed on the sets of items being considered, but items are only eliminated from consideration if they are clearly flawed. Bock et al. (1997) indicated that the 100 sampled spelling words differed substantially on item statistics such as difficulty estimates and measures of discrimination, but all were used for their test. For their example, only a single score was reported although different types of words clearly were included in the sample. Reporting was done in this manner because the goal is to estimate the total proportion of words from the full domain that can be spelled. This approach to test development is similar to "market-basket reporting," which is used to estimate changes in the cost of living. The costs of goods and services are obtained for a sample of purchases that is designed to represent the purchasing habits of the typical family in the United States.

A Comparison of the Two Approaches to Test Development

On the surface, the two approaches appear similar. Both select a set of test items to administer, and both are designed to estimate a single score. However, the surface view is deceptive. The continuum approach does not require a random sample of items. Instead, item selection is done in a way that allows accurate estimation of the location on a hypothetical continuum. If the goal is to estimate the level of a person's mechanical aptitude, the items are selected to cover a range of difficulty that will allow accurate location of the person on a mechanical aptitude scale. It is not clear how many items are in the full domain of possible mechanical aptitude items, and it is not necessary to know that. It is only necessary to be confident that the items are good indicators of location on the scale. This approach parallels the example of finding the person on K Street. The continuum approach is consistent with IRT and CAT. Much of the technology used for the analysis of tests is based on the continuum model.

The domain approach is based on the idea that it is possible to define the domain of interest and that it is important to know the percent of the entire domain that can be successfully performed by each individual. For example, arithmetic computation items that are the sum of two, two-digit positive integers are a well-defined domain consisting of 8,100 items. It would be impractical to administer all 8,100 items to students to determine the percent they can answer correctly, so it would be reasonable to randomly sample 20 of them to get an estimate of the total percent.

These simple domain examples seem to imply a single continuum, but the usual applications are more complex. Achievement tests are usually designed to get an estimate of the amount of the target curriculum for a subject matter area that has been acquired. Rather than these simple addition items, the curriculum may include addition, subtraction, multiplication, and division applied to simple computation problems and more realistic story problems. The problems might also contain integers, decimal numbers, and fractions as well. This domain is very complex and it is likely very challenging to determine the total number of items that make up the domain and the proportion that would fall into different stratification cells.

Even though the target domain for an achievement test is typically very complex (imagine one for 8th grade social studies), the scoring model that is used to get the estimate of the full domain that has been acquired is not. In many cases, summed scores, or proportion correct scores, are produced, or a closely related method, the Rasch model (Rasch, 1960), is used to estimate a score on the domain. These models treat all items that are scored 0 or 1 as equally informative, and the Rasch model assumes that all items are measuring a single dimension. Most achievement domains do not meet the requirements for the use of unidimensional psychometric models.

It is interesting that the proportion or percent-correct score does not require a unidimensional assumption if it is clear what the results means. Suppose I am having a garage sale with all kinds of objects: children's books, furniture, electronic equipment, clothing, and the like. After the sale is done, someone asks what proportion of the objects sold. If I had a count of the total number of objects, my answer might be that 50% of everything sold. But, that answer treats each object as one thing and ignores the difference in value. It may be that all the children's books have sold, but none of the furniture. The sales may not represent the full domain, but persons might assume that they do. The challenge to the domain approach is to get the representative sample of the domain and use that as the basis for reporting the results. But often the full domain is not specified in much detail. Instead, a set of items is produced and it is assumed that they are a representative sample of the full domain. The full domain is the collection of items, that when randomly sampled, would give the set that of items that was actually used on the test. To the extent that the assumption is not true, the results will be misinterpreted—thinking that half of the furniture sold at the garage sale when in fact none did.

The most critical issue with the domain model is that the analysis technology that has been developed based on the continuum model is not a good match to tests produced assuming the domain model. At this time, that technology is dominated by methods that use IRT. Most achievement testing programs use IRT-based methods for item analysis, item selection, score reporting, and test score equating. Some use the IRT methods as the basis for CAT. Almost all the IRT models used assume a single hypothetical continuum for the representation of examinee performance. The more complex the domain is, the less the usually used IRT models match the complexity of the domain. This mismatch of analysis procedures and domain characteristics may not be too serious of an issue if the goal is to get an approximation of the average of all the components to the domain. However, users of test results usually want much more information from an achievement test than this simple approximation. The next section of this report describes what users of an achievement testing program want from it and how the domain and continuum models are related to those desires.

Desired Interpretations of Achievement Assessment Results

A cursory review of the types of results reported from state achievement testing programs[1] and the ways that results are discussed in popular newspaper articles[2] will show that there are at least three common desired interpretations of achievement assessment results. The first of these has become very popular in recent years as more focus has been placed on holding schools and teachers accountable. Policy makers have a strong desire for information about student growth in achievement from the assessment program. This indicator of growth is usually an indicator of change from one grade level to the next. This type of use of the results of achievement assessment programs is most consistent with the continuum model. This use has implied assumptions that students can be located as a point on a continuum, the continuum is the same across grades, and the change in location is an indicator of growth in achievement.

A second desired interpretation is that the content of instruction for an academic year is complex, and it would be useful to know how students are doing on the various parts of the instructional content. School administrators often want subscores or diagnostic classifications that will give information about performance on the various parts of the instructional content. This way of interpreting the results of an assessment program is consistent with the domain model, especially if the sampling of the domain can be thought of as stratified random sampling. If there are many cells in the sampling plan, why not report results for each of the cells?

A third desired interpretation is that the test items used in the assessment yield good examples of the kinds of performance that students are expected to exhibit in the classroom (see Popham, Cruse, Rankin, Sandifer, & Williams, 1985, for example). That is, the test items are good examples of instructional tasks. If that is the case, then teaching to the test is the same thing as teaching to good instructional targets. This interpretation is also consistent with the domain sampling model. Educational policy makers expect that some of the tasks in the domain require higher order thinking and complex problem solving so the stratified sample from the domain should include items that elicit evidence of those types of behavior. And, teachers should also be helping students develop those skills, so including such items on the assessment will encourage teachers to include instruction to develop those skills. Developing tests as instructional targets tends to make them more multidimensional and less consistent with the continuum model. Complex items usually need multiple kinds of skills to arrive at a good response. This level of complexity is part of the reason for wanting to include them.

The fact that users of test results have more than one desired interpretation results in problems for the test development and analysis. The focus on growth is consistent with constructing the test to define a common continuum across grades. It is also consistent with the conceptual framework for unidimensional IRT models and much of the technology for equating and vertical scaling. But the desire for subscores and diagnostic classifications is consistent with a multidimensional view of the content of the test. The usual analogy for measuring growth is the set of marks made on a wall to show the change of height of children as they get older. This is an example of the continuum model. But if parents are interested in change in height, weight, strength, flexibility, achievement, and so on, the marks on the wall are not sufficient to capture all that is desired. Further, if one composite of physical attributes was desired to show the growth of children, it is not clear what the best one would be. Certainly, the average of all the things being considered would not make sense. So it is when growth in achievement is considered as growth in a complex domain of moderately related skills and knowledge.

The desire to create test items that are good targets for instruction creates more problems. Such items tend to use multiple skills and knowledge that are difficult to separate. Such items also tend to take significant amounts of testing time so they reduce the opportunities to get good domain coverage.

The fanciful example of finding the location of a person on K Street might help clarify the challenges. If we are trying to track someone's movement on K Street, we can go through the question asking process twice and see how the estimate of location has changed. Posing these two sets of questions can be done efficiently. It is the equivalent of assessing growth.

But, if we restrict the development of the questions that are asked, requiring that they must be a representative sample of all locations in Washington DC, the set of questions will have to include many that are uninformative about the person's location (i.e., asking about the Smithsonian Castle when the person is near Rock Creek Park), and a two-dimensional model will be needed instead of a simple unidimensional one. Alternatively, if there really is interest in the location of the person in DC (not just K Street), forcing a unidimensional analysis model is like only being allowed to ask questions about K Street when the person is near the National Cathedral. In the first case, the process is very inefficient. In the second one, the constraints will make it impossible to accurately locate the person.

Having It All

Acknowledging the problems and reconsidering them as challenges could potentially result in an assessment system that would support all the interpretations that are desired. For example, if both growth in performance and assessment of subdomains are desired, a testing program that accepts the multidimensionality of the domain could support both interpretations. One way is to identify the subscores that are desired and develop short CATs for each of them. Then, a composite score from the subscores could be defined based on a rational analysis of the domain and the defined composite could be used as the growth measure. This process is essentially how the Armed Services Vocational Aptitude Battery (ASVAB) is developed (see Carretta, 2014, for one of many papers on the topic of using composites of subscores). Short adaptive tests are developed for many test areas, and then each military service forms a composite of the resulting scores that best predicts performance in training schools for that service (see Segall & Moreno, 1999, for the details of ASVAB development). A parallel approach could be adopted for achievement tests.

Another approach would be to accept the multidimensionality and use a multidimensional IRT (MIRT) model instead of the unidimensional model. Then, specific directions in the multidimensional solution could be selected to specify subscores to be reported as well as a composite scale for the purposes of tracking overall growth (see Reckase, 2009, for examples of this approach). This approach is more challenging than the first because MIRT linking and equating have not yet been used for large-scale testing programs. However, the approach has been demonstrated in research papers (see Reckase & Xu, 2015, for one example). This approach would take more initial development to determine solutions to practical implementation problems.

A third approach might be to collect the actual outcomes of instructional activities and evaluate those as the assessment. Work has been done in the past to develop paper-and-pencil portfolios as structured collections of the students' work. These portfolios were evaluated through rigorous scoring processes based on formal scoring rubrics (Reckase, 1995). With greater movement of computers into the classroom, the collection of work could be done electronically, and the scoring could be done by combinations of human and automated processes.

A fourth approach is more at a stage of conceptual development than the previous three. Complex test items or simulation tasks could be developed that, in the aggregate, require all the skills and knowledge that are in the target domain. With computer technology, a complete record of the work that students do when interacting with the task can be collected. This record can include every keystroke and the amount of time everything takes. It might even include eye tracking. To make this approach work as desired, models of learning and students' interactions with complex tasks would need to be developed and evaluated. Current work in this area is in its infancy (one example is given by Cohen & Kolstad, April, 2000). But, with more resources, this approach may be shown to be viable.

Although work has been done on these alternatives, little interest has been demonstrated in using them for large-scale achievement testing. They seem too complicated, or there has been no demonstration that they will work in practical settings. Also, most achievement testing programs do not have the resources or time available to do the research and development needed to make the last three of the four alternatives described into practical methods. So, the current state of affairs leads to testing programs that have ambiguous interpretations.

The current state of large-scale achievement testing can be summarized as follows:

1. Develop the test using a domain sampling model. Specify the domain through a description of the requirements for the curriculum and develop a test blueprint that is consistent with the idea of stratified sampling, although actual sampling from a domain is seldom performed.
2. Include in the domain items that are simulations or "big" items that get at multiple skills and levels of cognitive complexity. These items are scored either using overall ratings or as "analytic scoring" that produces scores for parts. These items are sometimes treated as testlets to account for the dependencies in the parts.
3. Even though the first two descriptions yield tests that measure many skills and abilities, use the technology based on the continuum model to analyze the data. This approach creates a composite score for reporting, but the composite is defined by the internal workings of the analysis program rather than by a rational analysis of the domain. Also, the composite may shift from grade to grade, adding error to growth estimates.

The current approach to large-scale achievement testing has other problems that are symptoms of the conflict between the two development models. Items are sometimes deleted because of lack of fit to the unidimensional IRT model. These deletions may reduce domain coverage in favor of matching the continuum model. The opposite effect is adding content balancing to a CAT to deal with the complexity of the domain even though a unidimensional model is assumed for operating the CAT. If the set of item response data can be well fit by a unidimensional model, then content balancing should not be necessary. All the items are measuring the same construct. Testing programs often go even further. The unidimensional model is considered well supported by the data, but then there is a desire to report subscores implying that the response data need a multidimensional model for accurate representation.

The result of mixing these two development concepts is confusing interpretations and continuing attempts to report results that are weakly supported by the models. By emphasizing the inconsistencies between the two models, my hope is that greater clarity can be achieved when tests are developed.

What Can Be Done to Improve the Situation?

It is easy to criticize and complain. It is challenging to arrive at solutions. One approach to improve the design of tests is to acknowledge the multiple uses that are desired for test results and the limited resources available to support all of those uses. One way to do that is to start the process by developing a list of inferences that users want to make from the results of the test. Ultimately, each of the inferences should be supported with a validity argument based on the data obtained from the testing program. The list of desired inferences can then be prioritized in importance. If an inference about the growth in student achievement is most important, then the development of a test based on the continuum model needs to be emphasized. If an inference about the coverage of the domain is most important, then a good domain description and a sampling plan is needed. Each of the desired inferences has a cost. Adding more inferences have additional costs, and it costs more if the models are inconsistent. Developers can start at the top of the list and work down until the resources are used up. Someone will then have the difficult task of telling users that those inferences that are below that point on the list cannot be supported.

This approach has the advantage that some parts of the validity arguments for the desired inferences are developed as a natural part of the test design and development process. In the past, validity studies tended to be performed after a testing program was operational rather than at the beginning. There is also a tendency to add desired inferences after the test development is done. I expect there is no way to avoid wanting to add uses, but it would be helpful if the validity argument was the first step when any additional inferences are considered. However, when resources run out, someone must say it is not possible to add these additional inferences to the interpretation of the results of the program.

I must confess that I have been guilty of doing the things I am now arguing against. As a test designer and developer, I have often tried to give users everything they asked for even if it did not seem like a good idea. I know users want subscores even if they do not know what they will do with them. I know that growth measures are desired even when there are changes in the curriculum from grade to grade. I know policy makers want the test to cover everything to make sure everything is taught. The only way to do all of this is to get more information from the testing program, but policy makers also want to reduce testing time. One solution is to collect all the work from students throughout the academic year and figure out a way to score the results of each assignment. Then it would also be required that the results from all the assignments could be used to estimate locations on scales that are meaningful for reporting results. Then of course, information would be required to support inferences that are drawn from these results. Teachers already do these tasks to some extent. They compile all the information to report grades. But the grades are not comparable across classrooms, so it is not possible to draw all the inferences that are desired by policy makers. The typical result is a grade point average that has uncertain meaning.

Developing models to analyze the collected information from a year of instruction will not be easy. It will likely take years of research to provide information to support all the inferences that users want. But, I am optimistic. I think it can be done. The challenges will be how to satisfy those who want everything today and how to get more information per unit of time so that all the inferences can be supported.

Final Thoughts

The major factor that limits our capabilities to get information to support all the desired inferences is time. It is difficult to get sufficient information for accurate estimates of location or proportion of a domain when there is only time for the administration of a relatively small number of dichotomously, or even polytomously, scored items. Sixty test questions seem like a lot, and administration of them takes at least an hour, and certainly more time if they are complex items. But, the kind of picture of students' skills and knowledge that can be obtained using about 60 pieces of information is very limited.

To make this point, consider the picture of a student (Figure 2) that was made with 56 integers represented by gray pixels. The integer numbers were turned into shades of gray to generate the picture. This amount of information does not provide much of a picture. Even if more levels of information were obtained from each item to allow a color image to be produced, the picture is still not very informative (see Figure 3).

A good picture requires millions of pieces of information, 7,990,272 pixels in this case (see Figure 4). This need for more information is the problem that we encounter when trying to get an accurate representation of what students know and can do. With limited time, the result is a very low-resolution picture. The hope is that with more

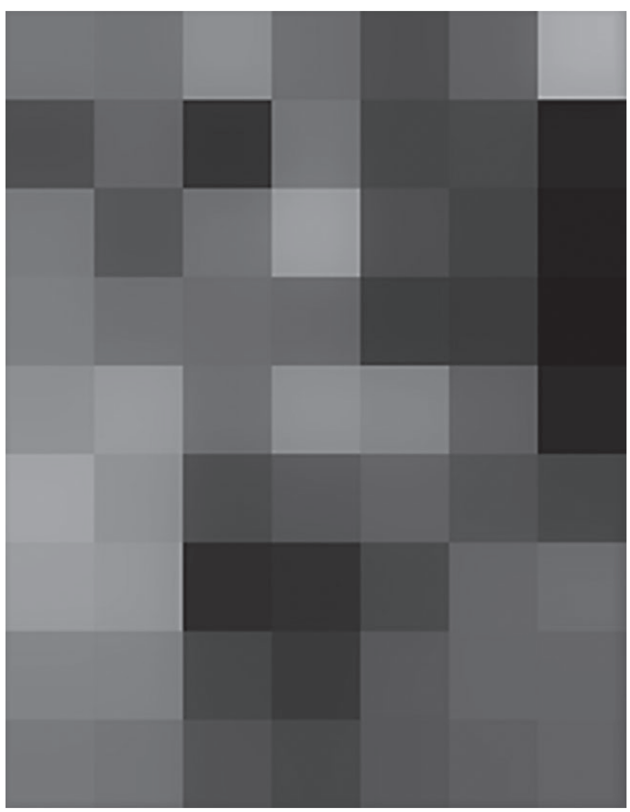


Figure 2 Picture represented with 56 gray pixels.

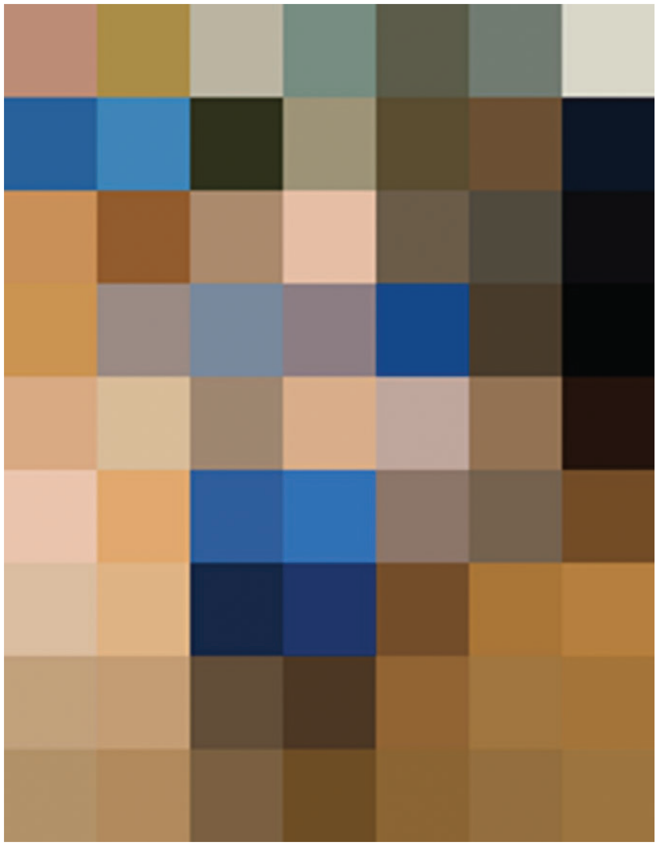


Figure 3 Picture represented with 56 colored pixels.

research we can collect more useful information using technology so that we can give users the level of detail that they want.

Do Achievement Assessments Result in Achievement Measurements?

In my opinion, the answer is no. One of two things tends to happen. Either the continuum model is distorted by forcing domain sampling on it, violating the assumption of a single continuum, or the domain model is well implemented, but then the data are analyzed using the technology developed for the continuum model. In the former case, the estimates of location on the continuum are tainted by information that is unrelated to it. In the latter case, the information from the domain is projected onto an ill-defined composite of all the concepts that are included in the domain. In either case, the meaning of the reported score is not clear.

The purpose of this paper has been to highlight conflicting uses of models that cause problems in the interpretation and use of test scores. In some cases, the results from using both conceptual models are not too bad. The domain that is sampled may have a hierarchical content structure that can be reasonably approximated by a continuum. However, in general, that is not the case.

Ultimately, those who develop tests and report results should seek to do no harm. They should try to make sure that what is reported can be interpreted accurately. It is doubtful that scores on individual items or highly correlated subscores can be interpreted in a meaningful way. If the domain model is used, care should be taken that the test really is a representative sample of the domain so the scores have an interpretation that can be supported.

It seems appropriate to finish with a quote from those great students of testing policy, Keith Richards and Mick Jagger: "You can't always get what you want, but if you try sometimes well you might find, you get what you need" (Jagger & Richards, 1969, side 2, band 4). Although it is not possible to give users and policy makers everything they want, maybe we can give them what they need.

Figure 4 High resolution picture.

References

ACT, Inc. (2006). *COMPASS/ESL reference manual*. Iowa City, IA: Author.

American College Testing Program. (1976). *ACT proficiency examination program (PEP)*. Iowa City, IA: Author.

American College Testing Program. (1989). *Preliminary technical manual for the enhanced ACT assessment*. Iowa City, IA: Author.

Bock, R. D., Thissen, D., & Zimowski, M. F. (1997). IRT estimation of domain scores. *Journal of Educational Measurement, 34*(3), 197–211.

Brown, W. (1911). *The essentials of mental measurement*. Cambridge, England: Cambridge University Press.

Carretta, T. R. (2014). *Predictive validity for the Armed Services Vocational Aptitude Battery for several US Air Force enlisted training specialties* (Technical Report No. AFHRL-RH-WP-TP-2014-0046). Dayton, OH: Wright-Patterson Air Force Base, Air Force Research Laboratory.

Cohen, J., & Kolstad, A. (April, 2000). *Theory-consistent item response models*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Cornman, O. P. (1902). *Spelling in the elementary school: An experimental and statistical investigation*. Boston, MA: Ginn & Company.

Downing, S. M. (2006). Twelve steps for effective test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development*. Mahwah, NJ: Lawrence Erlbaum.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.

Greene, H. A., Jorgensen, A. N., & Gerberich, J. R. (1944). *Measurement and evaluation in the secondary school*. New York, NY: Longmans, Green and Co.

Hardy, G. H. (1960). *A course of pure mathematics* (10th ed.). London, England: Cambridge University Press.

Jagger, M., & Richards, K. (1969). You can't always get what you want (Recorded by The Rolling Stones). On *Let It Bleed* (LP). New York, NY: London Records.

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement: Vol. I. Additive and polynomial representations*. New York, NY: Academic Press.

Leslie, L. A., & Funk, C. E. (1935). *25,000 words spelled, divided, and accented*. New York, NY: Funk & Wagnalls.

Luce, R. D., Krantz, D. H., Suppes, P., & Tversky, A. (1990). *Foundations of measurement: Vol. 3. Representation, axiomatization, and invariance*. San Diego, CA: Academic Press.

Mehrens, W. A., & Lehmann, I. J. (1991). *Measurement and evaluation in education and psychology* (4th ed.). Fort Worth, TX: Holt, Rinehart and Winston.

Monge, R. H., & Gardner, E. F. (1970). Syracuse University adult development study: A progress report. In W. S. Griffith & A. P. Hays (Eds.), *Adult basic education: The state of the art*. Washington DC: Government Printing Office.

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning*. Urbana: University of Illinois Press.

Popham, W. J., Cruse, K. L., Rankin, S. C., Sandifer, P. D., & Williams, P. L. (1985). Measurement-driven instruction: It's on the road. *The Phi Delta Kappan, 66*(9), 628–634.

Porter, A. C. (2002). Measuring the content of instruction: Uses in research and practice. *Educational Research, 31*(7), 3–14.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research.

Reckase, M. D. (1979). Unifactor latent trait models applied to the multifactor tests: Results and implications. *Journal of Educational Statistics, 4*(3), 207–230.

Reckase, M. D. (1995). Practical experiences in implementing a national portfolio model at the high school level. *NASSP Bulletin, 79*(573), 31–36.

Reckase, M. D. (2009). *Multidimensional item response theory*. New York, NY: Springer.

Reckase, M. D., & Xu, J.-R. (2015). The evidence of a subscore structure in a test of English language competency for English language learners. *Educational and Psychological Measurement, 75*(5), 805–825.

Rice, J. M. (1897). The futility of the spelling grind. *Forum, 23,* 163–172, 409–419.

Segall, D. O., & Moreno, K. E. (1999). Development of the computerized adaptive testing version of the Armed Services Vocational Aptitude Battery. In F. Drasgow & J. B. Olson-Buchanan (Eds.), *Innovations in computerized assessment*. New York, NY: Psychology Press.

Suppes, P., Krantz, D. H., Luce, R. D., & Tversky, A. (1989). *Foundations of measurement: Vol. II. Geometrical, threshold, and probabilistic representations*. San Diego, CA: Academic Press.

TenBrink, T. D. (1974). *Evaluation: A practical guide for teachers*. New York, NY: McGraw-Hill.

Thorndike, E. L. (1904). *An introduction to the theory of mental and social measurements*. New York, NY: The Science Press.

Wissler, C. (1901). The correlation of mental and physical tests [monograph supplement]. *Psychological Review, 3*(6), i–62.

Notes

1 The departments of education for states in the United States have websites that provide a lot of information about the state testing system and the uses of scores. One example with extensive information is the Georgia Department of Education (schoolgrades.georgia.gov and www.gadoe.org).

2 The *Los Angeles Times* has an archive of articles on achievement tests at this site: http://articles.latimes.com/keyword/achievement-tests.

Suggested citation:

Reckase, M. D. (2017). *A tale of two models: Sources of confusion in achievement testing* (Research Report No. RR-17-44). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12171

Action Editor: James Carlson

Reviewers: Michael Kane and Michael Zieky