Managing Multiple Measures

A white paper from MetaMetrics®, Inc.

by Gary L. Williamson, Ph.D., former Senior Research Associate

LEXILE® QUANTILE®

Table of Contents

Why We Have Multiple Measures

When the No Child Left Behind (NCLB) Act of 2001 was signed into law in January 2002, it established sweeping requirements related to annual achievement testing for state accountability purposes. Since that time, states have re-evaluated their testing programs, and in some cases expanded them, to ensure they meet the requirements of the law. However, states and school districts use achievement tests for purposes other than accountability (e.g., for instructional or programmatic monitoring) and often supplement the annual accountability assessments required by NCLB with interim tests during the school year to gauge whether they are on track to meet the annual achievement targets required by the law. Consequently, students are more likely than ever before to be assessed multiple times during a school year.

Multiple measures of a student's performance are also sometimes necessary to meet the professional standards jointly developed by the American Educational Research Association (AERA), the American Psychological Association (APA) and the National Council for Measurement in Education (NCME). Several of the standards (e.g., Standards 10.12, 11.20 and 13.7) enjoin educational professionals to consider collateral information when interpreting test scores and to avoid basing important decisions about a student on a single test score (Standards, 1999).

School districts design their testing programs to meet a variety of needs that include institutional accountability (e.g., for schools), individual student accountability (e.g., grade promotion standards), programmatic evaluation, instructional monitoring, progress monitoring, diagnosis of individual student learning needs and selection of students for special services or recognition, among others. Schools also increasingly use supplemental metrics such as The Lexile Framework® for Reading or The Quantile Framework® for Mathematics to enhance score interpretation and facilitate comparisons of performance derived from different reading or mathematics assessments. With increasing variety and frequency of assessments, combined with the inclination to employ a common metric for all assessments of a given construct, it is increasingly routine to have multiple, comparable assessment measures for reading or mathematics available for each student.

The Problem With Multiple Measures

Now that so much measurement is occurring, some educators are noticing that scores for the same individual differ on different occasions. This becomes more evident whenever a common scale is used for interpreting the scores from different tests. Thus, questions have arisen as to why differences occur and why they are (in some cases) so large.

[Note: It should be pointed out that, in this paper, we assume that a common developmental scale is being used to measure a given construct (e.g., reading or mathematics) even though different tests may be involved in the measurement on different occasions. When different tests are used to measure the same construct, comparisons of performance across occasions are facilitated by a common scale. However, as we shall see below, other factors can complicate those comparisons even when a common scale is used. That being the case, we will avoid discussion of situations where the scales differ. We will also avoid situations involving scales that are not equal-interval in their measurement characteristics. So, for example, we will not consider the use of achievement levels or performance levels as scales of measurement, even though they are widely used, because they are not equal-interval in nature.]

Williamson (2004) cited two fundamental reasons that an individual's scores may change from one occasion to another. The first reason is the uncertainty of measurement itself. That is, every measurement has some error associated with it. The second reason is that whenever substantial time has elapsed between measurements, growth may have taken place. Both influences can be operating simultaneously whenever data are collected for the same individual over long enough periods of time for growth to have occurred.

Historically, most teachers and parents have been oblivious to the uncertainty in the measurement of educational constructs. Most of our traditional assessments are either norm-referenced standardized tests such as the Stanford Achievement Test, 10th Edition (SAT–10), or high-stakes state assessments such as the Mississippi Curriculum Tests (MCT). In such cases, the tests are almost universally administered annually. Here, measurement error and growth are both present. However, when a test is only administered annually, in general, there is enough growth in the elapsed year that the amount of growth eclipses the amount of uncertainty in the measurements. Thus, the scores all seem to go in the right direction. That is, almost all students tend to show higher measures on the second assessment than on the first. Although the issue of measurement error rarely arises in these cases, it is still present.

An example from the measurement of a physical attribute (height) helps to illustrate this principle. Imagine taking a group of second graders and measuring their height in March 2005 and then measuring them again as third graders in March 2006. Practically everyone would be taller a year later as third graders. Imagine, however, that we measured their height as second graders on two consecutive days. If we rank-order the students on day one in terms of their height and rank-order them again on day two, the rank orderings would not be the same. Yet, it is not likely that substantial growth has taken place over just one day. We can imagine that some of the differences over two days could be attributed to wearing different shoes, the precision of our instrument (is a tape measure as accurate as a yardstick, etc.) or the time of day that the measure was taken (we are taller in the morning).

Likewise, if we gave the same norm-referenced or criterion-referenced test on two consecutive days, the scores for an individual would vary. One of the most widely talked about assessments in our country is the SAT from the College Board. In recognition of the uncertainty of measurement, college admission offices may accept the highest score from any administration. For example, if a student on day one scores 600 on verbal and 660 on math and on day two scores 660 on verbal and 600 on math, the student may report 660 on verbal and 660 on math.

In developing a strategy for managing multiple measures, it is useful to consider various influences that affect test scores under common conditions in education. Williamson (2004) described the nature of measurement error for individual scores when a single test has been administered on two occasions close together in time. This paper will briefly revisit that pic-

ture, and then expand the perspective to consider what happens when more than one test of the same construct has been administered in close temporal proximity, and then finally what happens when the timeframe is long enough that growth has taken place.

Why Scores Change

Measurement Error

From Williamson (2004), we recall that measurement error is variation in scores that is irrelevant to the construct being measured. This means that a student's score is influenced in part by factors other than the attribute that is being measured. Measurement error occurs with all measurements, whether they pertain to physical attributes (e.g., height) or psychological constructs (e.g., reading ability or mathematical understanding).

Measurement error can be random or systematic and can arise from sources both internal and external to the student. Williamson (2004) provided a table to illustrate sources of random errors of measurement. Random errors cause an individual's scores to be inconsistent from one administration of a test to another. For example, suppose a student is extremely hungry on one testing occasion and comfortably nourished on another. The scores may differ because of the difference in the student's physical state. This peculiarity would have no effect on any other student, just the individual who happened to miss breakfast that day.

As noted in Williamson (2004), a variety of random errors can potentially adversely affect the consistency of scores. Measurement error is the primary concern when examining scores from the same test on two successive occasions that are in close temporal proximity.

Random measurement errors (for measures in close temporal proximity) tend to be independent across people or occasions. So, averaging measures across people (for a group summary) or occasions (for an individual's score) is one strategy for managing the variability in scores that arises from random measurement errors. There are more and less sophisticated methods of averaging that can be used, depending on the context. Such methods are discussed further in the Strategies section.

Test-Specific Bias

Systematic errors of measurement are less frequent than random errors of measurement, but when they occur they affect the accuracy of measurement, preventing us from getting a true measure of a student's ability. They result in *bias*. One way in which this becomes particularly worrisome is when a student takes two different tests of the same construct, but the scores from one test are positively or negatively biased with respect to the student's true scores. Then the student's scores on the two measures will be different, but the difference is not primarily due to random influences on the measurement. This kind of situation sometimes occurs when one test is very important and has individual consequences for the student, while the other test does not.

For example, suppose a student takes a mathematics assessment at the end of the year and the student's promotion to the next grade will be determined in part by the test performance. It is likely that the student will study intensely and try very hard to score as high as possible on that test. Then suppose that the student is required to participate in another mathematics assessment that has no consequence for the student. It does not count toward the student's grade in the course, nor does the score get reported to anyone. It may be that the student will not put the same effort into the second test because of a lack of motivation to do so.

This scenario is not purely theoretical. Just such differences have been noted by psychometricians. Students who participate in item tryout studies (important to test construction) may not perform as well as comparable students who take the operational tests (where the scores count for them), even though both tests contain the same items (Thissen, 2005).

In some school districts, progress monitoring tests occur several times during the school year (e.g., fall, winter and spring), with a spring administration just weeks before or after the annual end-of-year high-stakes test. When the lower stakes progress monitoring test occurs after the high-stakes test, scores may appear to decline, causing concern for the teacher and parent. Yet, from the student's point of view, the progress-monitoring test may seem less important and not worth the same effort as the high-stakes test.

In such situations, reduced motivation can result in negatively biased (systematically lower) scores.

Other characteristics of tests or test usage can result in systematic differences in the scores produced from test administrations. For example, the mode of test administration might make a systematic difference when students are more familiar with one mode than another. This has been recognized as a challenge when trying to link or equate computer-administered tests with their paper-and-pencil counterparts (Eignor, 2005).

There are different implications for managing multiple measures, depending on whether score differences are due to measurement error or bias. As mentioned earlier, in the case of measurement error, some type of averaging can usually reduce the uncertainty in the observed scores. This is because measurement errors are random and tend to cancel each other out with enough replications of measurement. Unfortunately, in the case of bias, the effect persists through all replications of measurement, so it cannot be eliminated by averaging. We know neither its magnitude nor its direction (positive or negative) without some additional information. Dealing with bias requires additional information, usually derived from monitoring the test administration, interviewing the participants or engaging in a logical analysis. In reconciling score differences, sometimes it comes down to which score is most consistent with other available information (e.g., grades and student class work).

Differential Reliability of Tests

As discussed in Williamson (2004), consistency of measurement is called *reliability*, and psychometricians are concerned with producing tests that are highly reliable. This is manifested when alternate forms of the same test yield approximately the same relative ordering of individuals when administered on two occasions close in time. Reliability is quantified by an index that ranges from 0 to 1, with higher numbers indicating higher reliability. Within-grade alternate form reliabilities range from .75 to .85 for many tests used in common practice.

Since tests are never perfectly reliable (i.e., reliability equal to 1), it is useful to be able to characterize the amount of measurement error associated with scores. A more concrete way of conceptualizing the theoretical variability of scores is the *standard error of measurement (SEM),* a number that is expressed in the same metric used to report the scores.

Williamson (2004) presented a table that showed ranges of the SEM for selected reading tests, by grade. The SEMs ranged from 72L (72 Lexiles®) to 153L. From psychometric theory, we know that we can be very (95 percent) confident that, with repeated testing, observed scores will fall in a band that is four SEMs wide. However, even for very reliable tests (e.g., SEM = 72L), this band might be large (i.e., 288L wide). For less reliable tests, we might expect to see observed score variation that exceeds 600L on retesting.

There is an inverse mathematical relationship between reliability and the SEM. The higher the reliability, the lower the SEM is. This makes sense because higher reliability means scores are more consistent on retesting. If scores are more consistent, then they vary less, and consequently the SEM is smaller.

Knowledge of the reliability and SEM for a test facilitates interpretation of the scores obtained by students when the test is administered on one or more occasions. However, the situation becomes more complicated when two different tests of the same construct have been used, and they have different reliabilities and SEMs. This means one test result is more trustworthy (in the sense of being consistent or reliable) than the other. We might want to use such information in any strategy to understand the multiple measures.

One way in which differential reliabilities arise is in tests designed for different purposes. Suppose one test is highly targeted to the test-taking population, tailors its difficulty level to the pool of person abilities and consequently can reflect a very wide range of content coverage (say, in terms of reading material) to obtain optimal individual measurement for all of the students. [Examples of such tests might include review tests for a textbook or tests such as the Scholastic Reading Inventory–Interactive.] Contrast this to another test of the same construct,

but one designed to reflect grade-level content, that covers a narrower range of content and targets its difficulty level to the average individual in the population. [Examples of this kind of test might include most state tests used for evaluation or accountability.]

Even if these two tests have been calibrated to the same scale, and both are administered to the same population under the same conditions (e.g., high stakes or low stakes), the results are likely to be markedly different because of differences in the SEMs. The second test will measure better than the first (i.e., have a smaller SEM) for individuals at or near the population average, but the first test will measure better than the second test for those who are definitely above or below average. Although we may accrue the interpretive benefits of a common scale for scores on each test, we have to take into account other differences in test design, usage and purpose as we attempt to understand and manage multiple measures.

As mentioned in earlier sections, different strategies might be used for averaging different observed scores to obtain a more reasonable estimate of an individual's score. This section leads to the consideration of a weighted average whenever the observed scores come from tests that differ in their reliabilities. For example, we might want to take the reliability or the SEM into account in the averaging. Such a strategy will be considered further in the Strategies section.

Individual Growth

When substantial time has elapsed between measurements, growth probably has occurred and will eclipse the amount of measurement error in the scores on each occasion. The measurement error is still there but is just not noticed because the growth is more obvious. Whenever the same test, or an equivalent form, has been administered repeatedly for the purpose of measuring growth, we may think about the developmental trajectory or *growth curve* for the individual over time (Rogosa, Brandt and Zimowski, 1982). Such a framework allows for modeling the growth and measurement error simultaneously and also enables discussions about normal or expected growth (Williamson, 2006).

A Framework for Thinking About Multiple Measures of a Single Construct

Given the previous discussion, we might organize our thoughts about multiple measures into a framework that will help us develop strategies for managing multiple measures. We have seen that the time duration between measurements can be important because it helps us decide whether to focus primarily on measurement error, or whether we need to also consider growth. We have also seen that whether we use the same test for multiple measurements or whether we use different tests is important because that helps us decide whether we may need to consider differential reliabilities, or differences in purpose or usage context for the different tests.

We summarize these two dimensions (time duration and number of tests) into a table (see Figure 1). In the table, the two columns under "Time Duration" correspond to shorter or longer timeframes, and the two rows correspond to situations where multiple measures are derived from either the same test or different tests, respectively. Throughout the table we are

FIGURE 1: Primary Sources of Score Differences Under Different Testing Scenarios

Multiple Measures Derived From	Time Duration: Shorter	Time Duration: Longer
Same Test	Measurement Error	Measurement Error Reliability Growth
Different Tests (Same Construct and Scale)	Measurement Error Differential Bias Differential Reliability	Measurement Error Differential Bias Differential Reliability Growth

assuming that the same construct is being measured with a common scale under all of the conditions.

Some comment is appropriate regarding what we mean by "shorter" and "longer" times. Generally, when we think of shorter times, we are thinking about situations where tests have been administered to the same individuals within just a few weeks—say two to four weeks. In some cases, we may even consider several months to constitute "shorter" timeframes. More precisely, we are contemplating timeframes short enough that measurement error could eclipse any growth that may have occurred. By "longer" time duration, we have in mind a timeframe spanning perhaps a year or more. More precisely, these are timeframes long enough that growth must be considered in addition to any measurement error that is present in the scores.

In the table, we indicate the various factors discussed earlier that may influence why scores change under each combination of conditions related to time or number of tests used. So, for example, when the timeframe is very short and the same test has been used to produce multiple measurements, our primary concern is measurement error. Thus, in this case, any strategy for managing multiple measures must address the issue of measurement error.

When the timeframe is longer but the same test has been used consistently, we must not only consider measurement error but also growth. Because our ability to detect growth over time may depend on the reliability of the test (Rogosa and Willett, 1983), reliability is also listed as a possible influencing factor.

When the timeframe is short but different tests have been used, we see in the table that we must consider the influences of measurement error and differential reliabilities, and we may also need to consider possible sources of bias in one or both of the tests.

Finally, when the timeframe is long and different tests are in use, we have the most complex situation. In this case, measurement error, differential reliabilities, differential biases and growth may all be involved. Thus, any strategy to manage multiple measures in this setting must address all four possible influences.

Strategies for Understanding and Managing Multiple Measures

Scores in the Short Term (Weeks or Months)

Situation 1. Any time we attempt to measure a construct, we are obtaining an estimate. However, this estimate reflects our *uncertainty* (about the construct), which arises because of measurement error. That is, if we repeatedly measured an individual multiple times over a short time interval, the scores would not all be the same. We are uncertain which score is the best one to use. This fact holds whether we measure a construct such as blood pressure or reading ability or mathematical understanding.

The more measures of a construct we obtain, the more confident we can become in estimating the *true score*. Just as we can be more confident in estimating one's blood pressure if we take 10 measures spread over a few days, as opposed to one reading in the doctor's office, likewise we can place greater confidence in the inferences we draw about a student's reading ability or mathematical understanding if we can assemble 10 estimates (test scores) spread over a reasonable time period.

A natural question arises whenever we have collected multiple measures. How do we use the information to improve our confidence in our estimate of a person's true ability? When the same test has been used on two (or more) occasions in close temporal proximity, a simple way to deal with measurement error is to average the multiple scores available. In theory, if enough replications of measurement were available, such an averaging procedure would yield the true score of the individual. That many replications are never available. In most cases, we will only have two or maybe three scores available. So, we may not get the true score by averaging, but we will obtain a better estimate than we would have from any of the single test scores alone.

Situation 2. When different tests have been used within a short timeframe, we need a strategy to address measurement error, differential bias and differential reliability. As discussed earlier, there are no averaging strategies to deal with bias, so let us put that aside for the moment and consider how to address measurement error and differential reliabilities together.

As mentioned earlier, we can use a weighted average of scores from different tests when we have information available that one test may be more reliable than another. In such a case, we want to give more weight to the more reliable test and less weight to the less reliable one. This can be accomplished by weighting each test score according to its SEM, for example. When the SEMs for the two tests are equal, the weighted average is the same as a simple arithmetic average. So, this strategy has the nice feature that it also works for the simpler case where the same test was used. Except for bias, a strategy that uses weighted averages of test scores with weights derived from the standard errors of the tests would work for all of the scenarios where the timeframe is short, whether the same test or different tests have been used. [Note: Reciprocals of the squares of the standard errors of the tests are the optimal weights for minimizing the standard error of the average.]

To deal with bias, however, we must have additional information about the test administration, the typical performance of the student and the conditions of student performance. Such information might be collected from the teacher, student or parent. It could be used to eliminate scores that may appear to be aberrant, so that they do not adversely affect the estimate of true performance produced by any averaging procedure.

Situation 3. Sometimes multiple scores occur over a slightly longer duration (say months) but are still relatively close together. In these cases, some small amount of growth may have occurred, but its magnitude may be on the same order as the amount of measurement error in the scores.

For these cases, there are more sophisticated techniques for estimating true scores than simple arithmetic averages or weighted averages. Advanced statistical methods are available to take advantage of other information (e.g., other estimates of the student's ability) and to incorporate this information along with the current scores to get a more precise measure of the student's likely true score. These *Bayesian* techniques use *prior* information in conjunction with current performance to make a prediction that incorporates both sources of data and produces an updated estimate that has less uncertainly associated with it than the available observed scores have.

Bayesian estimation, augmented by some mild assumptions about the amount of short-term growth that is typical, can be used to produce estimates of future performance and growth when the amounts of longitudinal data are insufficient for formal growth modeling.

Scores Over Longer Time Spans (Multi-Year)

Situation 4. When multiple measures have been generated from the same test over four or more occasions, then growth modeling is appropriate. A number of authors have described modern statistical procedures to analyze longitudinal data (e.g., Goldstein, 1979, 1995; Raudenbush and Bryk, 2002; Singer and Willett, 2003). These methods are empirically based (i.e., depend on data) and involve an interconnected series of steps.

Typically, one begins with exploratory analyses to discover general temporal patterns in the data. These might consist of plots of scores over time, smoothing the trend (non-parametrically or with simple linear regression) and examining the results across individuals in the population. These steps help the investigator to decide on an appropriate mathematical form for modeling the growth during the observed timeframe. These steps are followed by formal multilevel modeling of the individual growth curves as well as the variation in the growth parameters in a population of individuals.

Multilevel modeling is a powerful technique for studying growth in groups of individuals. When longitudinal data are available only for a single student, then multilevel modeling cannot be used. Simpler approximations (e.g., simple linear regression) can be used. However, when panel norms for growth (Williamson, 2006) are available for a suitable reference population, the information may be used to help understand individual growth in new contexts.

Situation 5. When multiple measures have been generated over four or more occasions for more than one test, then it is possible to investigate growth on multiple measures. (Again, we're assuming that all tests measure the same construct.) This involves an advanced statistical technique called latent growth curve analysis (Bollen and Curran, 2005; or, see Singer and Willet, 2003, for a review of existing liter-

ature). In such an analysis, it is assumed that serial observations on two different measures, both subject to measurement error, represent the underlying true growth in a construct over time. The analysis attempts to estimate the true growth, using the observed data and a theoretical model that specifies how the observed data relate to the underlying trait on each occasion.

Summary

The Lexile Framework for Reading has been linked to many reading tests. The number of tests linked to The Quantile Framework for Mathematics is growing. The availability of a common metric accompanied by the proliferation of testing in the United States has resulted in multiple measures being available for students. Educators, parents and students are noticing more frequently that these scores sometimes disagree even though they came from tests administered in close temporal proximity.

The availability of a common scale enables one to compare the results from two assessments to have a better understanding of a child's reading ability or mathematical understanding. As a practice, it is always better to have multiple observations to understand a student's true reading (or mathematics) ability. In addition, all assessments—national norm-referenced tests, state-level criterion-referenced tests, diagnostic achievement tests, progress monitoring assessments, etc.—have some inherent measurement error. If assessments are measuring the same construct (as is the case, for example, with the METROPOLITAN8 Reading Comprehension Test and the Scholastic Reading Inventory–Interactive), then we should expect the results to be consistent (within measurement error). However, we should never expect to replicate scores exactly from two assessments when there are differences in the stakes attached to the results (purpose for administering the assessment), the format (both in terms of how the items appear and how they are administered), the test-session environment and/or the test administration procedures.

Where differences exist for individual students, the test administrations should be reviewed and evaluated to determine which measure is more reflective of the individual student's true ability. Each individual measure should always be evaluated in terms of its relationship to the other measures to monitor a student's true growth.

References

Bollen, K. A., & Curran, P. J. (2005). *Latent curve models: A structural equation perspective*. Wiley Series in Applied Probability and Statistics. New York: Wiley.

Eignor, D. R. (2005, June). Linking scores derived under different modes of test administration. Paper presented at the ETS Conference, *Linking and Aligning Scores and Scales,* Princeton, N.J.

Goldstein, H. (1979). *The design and analysis of longitudinal studies.* New York: Academic Press.

Goldstein, H. (1995). *Multilevel statistical models*. (2nd ed). New York: John Wiley.

Public Law 107-110. The No Child Left Behind Act of 2001.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. (2nd ed.). Thousand Oaks, CA: Sage Publications.

Rogosa, D. R., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin,* 92, 726-748.

Rogosa, D. R.., and Willett, J. B. (1983). Demonstrating the reliability of the difference score in the measurement of change. *Journal of Educational Measurement,* 20, 335-343.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York: Oxford University Press.

Standards for Educational and Psychological Testing. (1999). Washington, D.C.: American Educational Research Association.

Thissen, D. (2005, June). Linking assessments based on aggregate reporting: Background and issues. Paper presented at the ETS Conference, *Linking and Aligning Scores and Scales,* Princeton, N.J.

Williamson, G. L. (2004). *Why do scores change?* Durham, N.C.: MetaMetrics Inc.

Williamson, G. L. (2006). *What is expected growth?* Durham, N.C.: MetaMetrics Inc.

About the Author

Gary L. Williamson, Ph.D., is a former senior research associate with MetaMetrics, Inc. With more than 30 years of experience in educational research on the academic, state and school district levels, Williamson's specialty is quantitative methodology encompassing psychometric, mathematical and statistical applications to educational data. He has written and spoken extensively on the subjects of educational assessment and accountability. Williamson earned both a doctorate of philosophy in mathematical methods for educational research and a master's of science in statistics from Stanford University. He also holds a master's of education in educational research and evaluation from The University of North Carolina at Greensboro, and a bachelor's of science in mathematics from The University of North Carolina at Chapel Hill.

MetaMetrics, Inc.
1000 Park Forty Plaza Drive, Suite 120
Durham, North Carolina 27713

Phone: 919–547–3400/1–888–LEXILES
Fax: 919–547–3401
Web site: www.Lexile.com