A Long-Term Study of the Stability of Item Parameter Estimates and Measurement Scales

G. Gage Kingsbury

Northwest Evaluation Association

Reprint requests to:

G. Gage Kingsbury

Northwest Evaluation Association

5885 SW Meadows Road, suite 200

Lake Oswego, OR 97035

Running Head:

Stability of Measurement Scales

Abstract

Much of the development that is done with Item Response Theory depends on the property of item parameter invariance and the stability of measurement scales. In particular, adaptive testing requires stable measurement scales as do accountability systems like No Child Left Behind. This study examined drift in item difficulty estimates for several thousand test questions associated with reading and mathematics measurement scales over a period of 22 years. Results showed very high correlations between item difficulty estimates from the time at which they where originally calibrated and the current calibration. Further, the average drift in item difficulty estimates was less than .01 standard deviations. Finally, the average impact of change in item difficulty estimates was less than the smallest reported difference on the score scale for two actual tests. While the findings of the study depend on the developmental processes used in creating the measurement scale, they indicate that an IRT scale can be stable enough to allow consistent measurement for more than two decades.

A Long-Term Study of the Stability of Item Parameter Estimates and Measurement Scales

Item Response Theory (IRT: Lord & Novick, 1968; Lord, 1980) derives part of its appeal from the fact that it allows the creation of measurement scales that are independent of the particular sample of individuals or test questions used to create the scales, and invariant when applied to particular groups of individuals within the population of interest. Lord and Novick (1968, pp 360) describe these properties in this manner

> "Because of its definition, the item characteristic function necessarily remains invariant from one group of examinees to the next, at least among those groups used in defining the complete latent space. This means that any parameter describing the item characteristic function is an invariant item parameter."

This invariance property is exceptionally valuable, because it provides us with the capacity to build measurement scales that can be expected to maintain their measurement characteristics, even though we modify test forms, or implement an adaptive test. The

most direct application of the invariance property is seen in the development of item banks using IRT (Vale, 1986; van der Linden, 1986).

In practice, IRT item parameter *estimates* will not be invariant. Estimates will vary due to a number of factors that have been researched fairly extensively in the past. These factors include sampling fluctuation (Swaminathan & Gifford, 1983), departures from unidimensionality (Bejar, 1980), and other characteristics of the calibration design such as item context (Yen, 1980).

To make matters more complex, the type of test used to create item parameter estimates and the algorithm used to compute the estimates will also influence the stability of the item parameter estimates (Jae-Chun Ban, et al, 2001). All of these factors that may affect the accuracy of item parameter estimates suggest that we should be cautious in relying on the invariance property of IRT in practical settings without verification.

As we add items to an item bank, these factors may cause long-term drift in item parameter estimates and trait level estimates for test takers. For instance, a small departure from unidimensionality may make a group of items that are being added to an item bank appear slightly easier than they actually are. This will probably have little impact in the first year that the items are used operationally. However, the use of these items to help calibrate new field test items may cause the new field-test items to have difficulty estimates that are slightly more biased cumulatively. Over the course of several years, this could cause the entire scale to drift, reducing our ability to make long-term statements about student performance.

Even though long-term scale stability is imperative to our ability to observe patterns of growth across time, few studies have examined the long-term stability of IRT

item parameter estimates. Two studies that have investigated the issue were conducted by Bock, Muraki, and Pfeiffenberger (1988) and by Sykes and Fitzpatrick (1992).

Bock, et al. (1988) investigated the stability of the item parameter estimates in the 3-parameter logistic IRT model from the College Board Physics Achievement Test over a period of ten years using an ANOVA design and looking for a two-way interaction between items and occasions. The authors found that there was a statistically significant drift in item difficulty across time. The authors interpreted the drift as being due to changes in physics instruction across the time period under investigation. The authors performed a similar analysis of the College Board English Achievement Test, and found no evidence of drift. Since the focus of this study was on the development of a statistical model to allow for drift, rather than on the drift itself, the authors did not discuss the impact of the observed drift on test scores.

Sykes and Fitzpatrick (1992) investigated the stability of 1-parameter logistic item parameter estimates for 285 items from a professional licensure test administered over a period of five years. This study found drift in item difficulty parameter estimates that was directional, with items being estimated to be more difficult across time. When the investigators examined the source of the drift, it did not seem to be associated with item position or item type. As in the previous study, the authors hypothesized that the change in difficulty estimates was associated with changes in curricular emphasis. Since the emphasis in this study was on the covariates of drift, the authors didn't discuss the magnitude of change in candidate scores that might be caused by drift in item difficulty estimates.

The current study extends this earlier work in several ways. First, it investigates stability of item parameter estimates in two large item banks used to measure achievement in mathematics and reading, rather than parameter estimates from a set of items used in a single test. Second, it uses measurement scales that have been designed to measure student growth across time. Third, it uses a longer elapsed time since initial calibration, ranging from 7 to 22 years. Fourth, it attempts to estimate the amount of impact that item parameter drift might have on student scores. Primary questions to be addressed in this study are

1) How much drift in item parameter estimates is seen in item calibrations separated by as much as 22 years?
2) Is the magnitude of item-parameter drift associated with the elapsed time since the original item calibration?
3) What impact does this observed drift have on trait level estimates for test takers?

The stability of measurement scales may be viewed through two lenses. Using the first lens, we may investigate whether individual items have changes in their difficulty estimates across time. If there is more change (drift) than expected due to sampling variability, we may identify this as a problem with the invariance assumption. Using the second lens, we may ask what impact any identified drift may have on the test scores from our assessments and what impact the drift may have on decisions that are

made as a result of the assessments. The questions asked in this study allow us to investigate the issue through both lenses.

It should be noted that this study is not an analysis of calibration procedures. It is fairly clear that procedures for calibration have improved over the 22 years encompassed by this study. The question this study addresses is whether calibration estimates change as a function of time, given the same calibration procedure.

The Measurement Scales

The measurement scales used in this study are the reading and mathematics scales developed by the Northwest Evaluation Association. These scales, known as the RIT scales, are associated with large item banks that are used to develop achievement tests for use in a variety of school districts. The one-parameter logistic (1PL) IRT model (Wright, 1977) was used to create and maintain the underlying measurement scales used with these banks. The RIT scales are linear transformations of the θ scale, originally defined with a mean of 200 and a standard deviation of 10. Over the course of time, the mean performance level and standard deviation of student performance has changed, but the relationship between the RIT scale and the skills needed to obtain a given RIT score have remained constant.

Since the 1970s, thousands of items have been added to these item banks. Each item has been connected to the original measurement scale through the use of IRT procedures and systematic measurement practices (Ingebo, 1997). Each item has been connected to the original measurement scale through the use of IRT procedures and systematic calibration design.

These measurement scales are used to develop adaptive tests and to measure individual student growth. Since both of these activities depend to a great extent on the item parameter estimates, it is crucial that the invariance assumption hold in this application. While some variability in item parameter estimates is expected, too much variability could cause growth measurement to be quite problematic. Growth is a difficult quantity to measure under the best conditions, so a stable scale is a prerequisite to maintaining accuracy in growth measures.

Method

Items

3091 mathematics items and 1728 reading items were administered to students from grades 2 to 10 in 10 school districts from 7 different states as a part of their districtwide assessment programs in the 1999-2000 school year. Any particular student took approximately 50 mathematics items and 40 reading items. All items were multiple-choice, with original item difficulty estimates that were obtained at least 7 years prior to the study. Approximately 320 mathematics test forms and 160 reading test forms were used in the study. Over 100,000 student test events were used for the study.

Tests

The items were administered within the context of an achievement level test (Kingsbury & Houser, 1997). An achievement level test is a paper-and-pencil test that has approximately seven different forms (levels) designed to differ in difficulty. Students are administered a particular form chosen for them individually based on past tests scores or

using scores from a routing test. This design is similar to a two-stage adaptive test (Lord, 1971) which uses past information in lieu of a first stage. The original difficulty estimates (described below) were used for test design and scoring.

Any individual student took approximately 50 mathematics items or approximately 40 reading items. Since different achievement level tests were used in the different school districts involved in this study, any individual item was seen by only a small sample of the students involved. The combination of all test forms across all school districts and grades resulted in the sparse data matrix that was used for calibrating all of the items in the study.

Original item difficulty estimates

The original IRT item difficulty estimates for all of these items were created between 1977 and 1993. The mean time between the original calibration and the new calibration was 16 years and 1 month. The original item difficulty estimates were obtained using a marginal maximum-likelihood calibration procedure (Houser, Hathaway, & Ingebo, 1978).

New item difficulty estimates

The new item difficulty estimates were created using the data collected in the 1999-2000 school year. Since few students took the same items and no student took a very large percentage of the items, the calibration procedure used was a procedure designed for use with adaptive tests and other sparse data structures (Houser, Kingsbury, & Harris, 1997).

This procedure was used because it is the direct analog of the marginal maximum-likelihood calibration procedure originally used to calibrate the items.

After elimination of items with very small samples, 2359 items were available for use in mathematics, and 1392 items were available in reading. Calibration sample sizes for these items ranged from 300 students to over 10,000 students. A minimum student sample size of 300 was established to correspond to the minimum sample size that was allowed in the original calibration procedure.

Analysis

While there are a variety of statistical tools available for identifying parameter estimate drift (see Donoghue & Isham, 1998), the use of the 1PL model simplifies matters substantially. In this study, simple differences between original and new difficulty estimates were used. Two analyses were conducted for each measurement scale in the study.

Scale drift analysis. The scale drift analysis included several aspects. First, correlations between the new and original item difficulty estimates were calculated and compared to correlations seen in other studies using the same measurement scales. Next, frequency distributions of the differences between the original item difficulty estimates and the new item difficulty estimates were calculated. These allowed the examination of the variability in parameter estimates. Bias and mean absolute differences were also calculated and compared to standard deviations of student performance to begin to identify the impact of parameter drift. Finally, item parameter estimate differences were examined as a function of the original calibration date to identify whether the elapsed

time between the two calibrations contributed to observed drift. This last analysis used a subset of the available items (2204 items in mathematics and 1253 items in reading) because some items were originally calibrated across several testing seasons.

Impact analysis. A second method of analyzing the effect of change in calibrations over time is to ask whether that change has a noticeable impact on students' scores. In this analysis two representative test forms used in the study were chosen as example tests. The two forms were middle-difficulty forms used in the fifth grade in a suburban school district in Indiana. For each of these forms two raw-score-to-RIT scoring tables were created, one using the original parameter estimates and one using the new item parameter estimates (for the 1PL IRT model, a particular number-correct score is associated with a single scale score, dependent only on the item parameter estimates). The two scoring tables were then compared to identify the maximum difference caused by using the new item parameter estimates. By comparing the scale scores obtained from the two sets of calibrations for a particular raw score, we can identify how much a particular student's test score would have changed as a result of the scale drift.

Results

Scale Drift Analysis

The observed correlations between the original and new item difficulties estimates were .967 in mathematics and .976 in reading. While these correlations are close to unity, it is useful to compare them to correlations obtained in other studies using the same measurement scales. Ingebo (1997) described a series of experiments from the 1970s in which multiple, concurrent samples were drawn to calibrate a set of items from these

scales, to identify the consistency of the calibrations. In those studies, the correlations of mathematics difficulty estimates across samples ranged from .95 to .99, and the correlations for reading items ranged from .96 to .99. The results from the current study mimic those from the studies in which the samples were drawn concurrently.

Although the correlations provide some evidence of stability, they don't provide information about the differences observed on an item-by-item level. Figures 1 and 2 show the frequency distributions of the differences observed subtracting the new calibration of item difficulty from the original calibration for each item. It can be seen from these figures that the distributions are fairly symmetric around a difference of zero. It can further be seen that few items have difficulty differences of more than ten RIT points (approximately one θ unit). The distribution of differences in mathematics has items reaching further from zero than reading, but it is useful to remember that the item sample in mathematics (N = 2359) is nearly twice as large as the sample in reading (N = 1392). Due to this discrepancy in sample sizes, we would expect to see more extreme differences in observing the mathematics items.

[Insert Figures 1 and 2 here]

Figures 3 and 4 show the relationship between the original item difficulty estimates and the new difficulty estimates for each item in each subject area. The relationships appear visually linear, and correspond well to the superimposed line of identity. The figures show no evidence of drift associated with the difficulty of the items, and give no indication of a non-linear trend in the item calibrations. It can be seen by a comparison of these two figures that the relationship between the original calibrations and the new calibrations for reading is slightly more consistent than that for mathematics.

Part of this visual difference is again due to the larger number of items in the mathematics comparison, but it is also due to the slightly higher correlation seen in the reading results.

[Insert Figures 3 and 4 here]

The directional drift (bias or average difference) in item difficulty estimates was -.11 RIT points in reading and -.17 RIT points in mathematics. To put this difference in context, the standard deviation of students' scores in sixth grade in the most recent norming study done using these measurement scales (NWEA, 2005) was 14.5 RIT points in reading and 14.8 RIT points in mathematics. The drift that has occurred in the scale over the 16.1 years of elapsed time in the studied interval has had an impact of approximately .01 standard deviations on the mean item difficulty estimate.

The average absolute difference in parameter estimates was 3.29 RIT points in reading and 4.53 RIT points in mathematics. The median absolute difference was 3 RIT points in reading and 4 points in mathematics. As expected, this difference was larger than the directional drift, but still less than one-third of a standard deviation. Given the small values for directional drift, we would expect these differences to balance out in a test of reasonable length. This assumption will be investigated more completely in the impact analysis below.

An additional question of interest is the relationship between the length of time since the original calibration and the difference in item difficulty estimates. Figures 5 and 6 show the difference in item difficulty estimates as a function of the initial calibration date. Two aspects of these figures are worth noticing. First, there was no noticeable directional impact of elapsed time on difficulty estimates. This indicates that

no easily observable directional drift took place in the scale values. Second, the variability of new difficulty estimates around initial estimates did not seem to vary systematically as a function of the time since original calibration. This indicates that the variability seen was a function of elements of the calibration design that is not influenced by the time since first calibration.

[Insert Figures 5 and 6 here]

Impact Analysis

Figures 7 and 8 show the RIT scores obtained from each number correct score for original and new difficulty estimates. The figures show clearly that the scores from the two different sets of item difficulty values are quite small. The differences in the two sets of scores are difficult to discern, since they are so similar.

[Insert Figures 7 and 8 here]

For the mathematics test, the maximum difference that could occur was 1.1 RIT points and the average magnitude of difference was less than .5 RIT points. For the reading test, the maximum valid difference was again 1.1 RIT points, and the average magnitude of difference was .7 RIT points. Since the smallest observable difference between two RIT scores is 1 RIT point, these differences are small enough to be rarely observable. Given a typical distribution of RIT scores, the difference would be less than 1 RIT point in 99 of 100 cases. Since the standard error of a score on one or these tests would be approximately 4 RIT points, the impact of the change in calibrations would not be expected to change instructional decisions.

Discussion and Conclusions

Two major conclusions from the study are as follows:

1) There was no substantial drift in item difficulty estimates across the timeframe of this study, and no trend was seen in changes in difficulty estimates as a function of time since initial calibration.
2) The largest observed change in student scores moving from the original calibrations to the new calibrations was 1.1 RIT points, with over 99% of expected changes being less that 1 RIT point.

While the overall conclusion of the study is that the measurement scales examined are stable across time, some individual items fluctuated noticeably from their original calibrations. This suggests the need for ongoing calibration analysis. Even with a fairly stable scale, individual items may have difficulties that vary across time. A follow on study will investigate the characteristics of these highly variable items. This study should enable us to identify whether the large changes in difficulty for a small number of items are possibly due to specific features of certain items or whether they might be due to changes in instruction that have reduced (or increased) a student's opportunity to learn the content in the question. Examples of items that might experience such fluctuation include the following:

- An example of change specific to the item would be an item asking for a definition of the word “radical” which has had three most common definitions since the early 1980s.
- An example of change related to opportunity to learn might be seen in an item asking about the characteristics of a retro-virus. In the late 1970’s only college-level biology students would have been introduced to the concept, but now it is standard content in most high school biology courses.

Building a stable measurement scale is as much an exercise in engineering as it is an exercise in calibration. The measurement scales under consideration here were originally designed using a four-square design (Wright, 1977) with multiple cross links within and across student grades. It is expected that this original development has contributed to the ongoing stability of the measurement scales studied here. Therefore, while this study has indicated that stable measurement scales can be created in practice, it does not suggest that the use of IRT calibration alone will assure scale stability.

In public education, there is ongoing debate about the quality of schools. One overlooked element that causes the debate to continue is the inconsistent nature of much achievement information. Different tests are used to measure student achievement in different grades in many locations. If there isn’t a consistent measurement scale linking these tests, comparison of performance across grades is difficult. In many cases, score equating is used to allow comparison from one year to another. While this is a useful statistical technique, it isn’t designed to create stable measurement scales.

A stable measurement scale allows the development of curriculum-referenced interpretation of test scores. For instance, with the mathematics RIT scale, a student who was able to complete two-digit addition question correctly would obtain the same score in the year 2002 that they would have obtained in 1980. With this development, changes in test scores can be related directly to changes in student capabilities. In turn, this will allow the identification of positive and negative trends in education as they happen.

The procedures used in creating the measurement scales examined in this study have been successful in creating stability. This is a requirement for good measurement and even more important if we are planning to measure change in a school or a nation across time. The results of this study indicate that we can create measurement scales that are meaningful not only for short-term comparisons, but for long-term studies as well.

References

Ban, J-C., Hanson, B. A., Wang, T., Yi, Q., & Harris, D. J. (2001). A Comparative Study of On-Line Pretest Item: Calibration/Scaling Methods in Computerized Adaptive Testing. *Journal of Educational Measurement*, *38*, 191-212.

Bejar, I. I. (1980). A procedure for investigating the unidimensionality of achievement tests based on item parameter estimates. *Journal of Educational Measurement*, *17*, 283-296.

Bock, R. D., Muraki, E., & Pfeiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement*, *25*, 275-285.

Donoghue, J. R. & Isham, S. P. (1998). A comparison of procedures to detect item parameter drift. *Applied Psychological Measurement*, *22*, 33-51.

Houser, R. L., Kingsbury, G. G., & Harris, G. (1997). MATCAL: A program for calibrating items using data from sparse matrices. Portland, OR: NWEA.

Houser, R. L., Hathaway, W. E., & Ingebo, G. S. (1983). An alternate procedure to obtain ability estimates in latent trait models. Paper presented to the annual meeting of the American Educational Research Association, Montreal, Canada.

Ingebo, G. S. (1997). Probability in the measure of achievement. Chicago, IL: MESA Press.

Kingsbury, G. G. & Houser, R. L. (1997). Using data from a level testing system to change a school district. In J. O'Reilly (Ed.), *The Rasch tiger ten years later:*

Using IRT techniques to measure achievement in schools (pp. 10 - 24). Chicago, IL: National Association of Test Directors.

Lord, F. M. (1971). A theoretical study of two-stage testing. *Psychometrika, 36,* 227-242.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

NWEA (2005). *RIT scale norms for use with Measures of Academic Progress and Achievement Level Tests*. Portland, OR: Author.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danske Paedagogiske Institut.

Swaminathan, H. & Gifford J. A. (1983). Estimation of parameters in the three parameter latent trait model. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing.* New York: Academic Press.

Sykes, R. C. & Fitzpatrick, A. R. (1992). The stability of IRT *b* values. *Journal of Educational Measurement, 29,* 201-211.

Vale, C. D. (1986**).** Linking Item Parameters Onto a Common Scale. *Applied Psychological Measurement, 10,* 333-344.

van der Linden, W. J. (1986**).** The changing conception of measurement in education and psychology. *Applied Psychological Measurement, 10,* 325-332.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14,* 97-116.

Yen, W. M. (1980). The extent, causes, and importance of context effects on item parameters for two latent trait models. *Journal of Educational Measurement, 17*, 297-311.

Tables and Figures

Figure 1.

Frequency of mathematics items as a function of the difference between the original item difficulty estimate and the new item difficulty estimate on the RIT scale (rounded to the nearest integer)

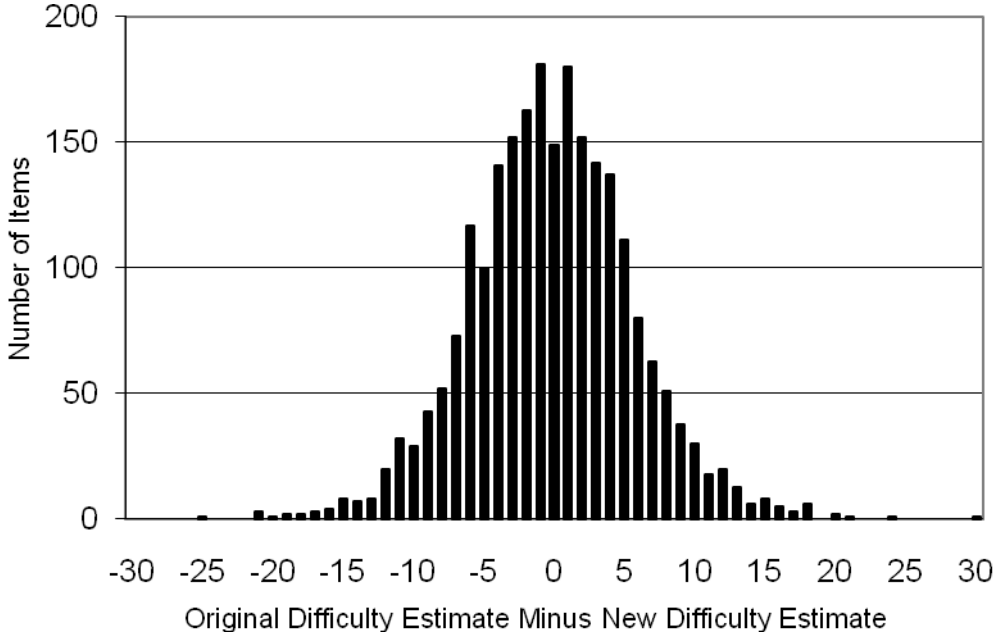


Figure 2.

Frequency of reading items as a function of the difference between the original item difficulty estimate and the new item difficulty estimate on the RIT scale (rounded to the nearest integer)

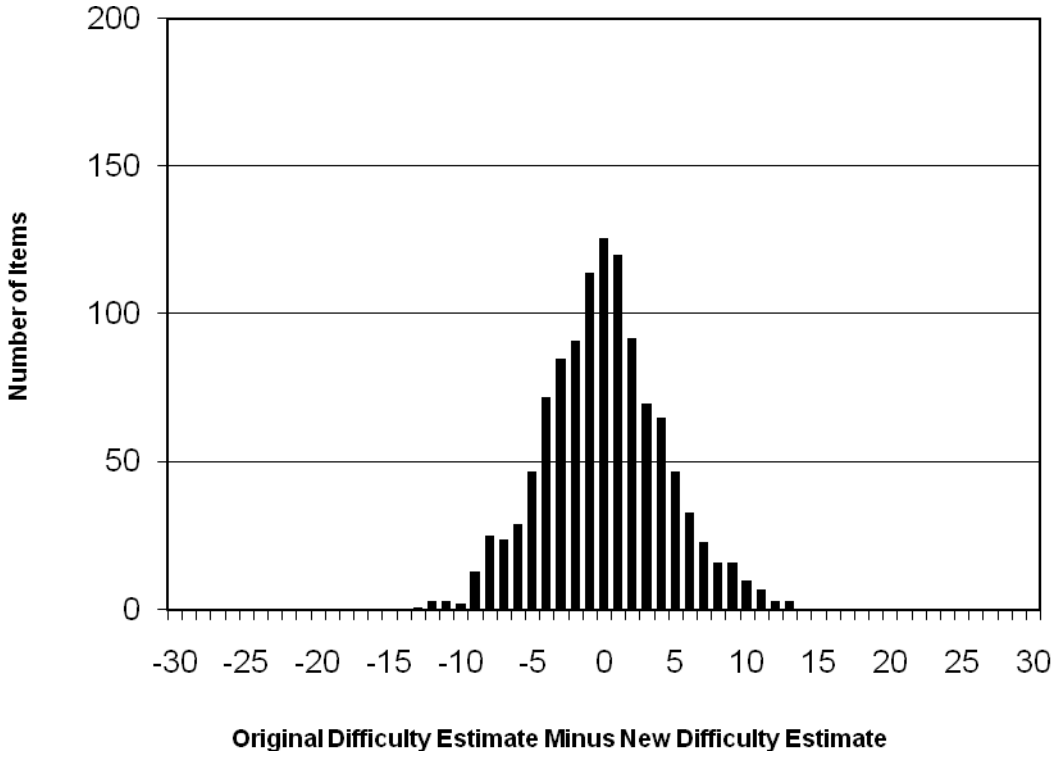


Figure 3.

Relationship of original and new item difficulty estimates on the RIT scale for 2359 items in mathematics with superimposed identity line (r = .967)

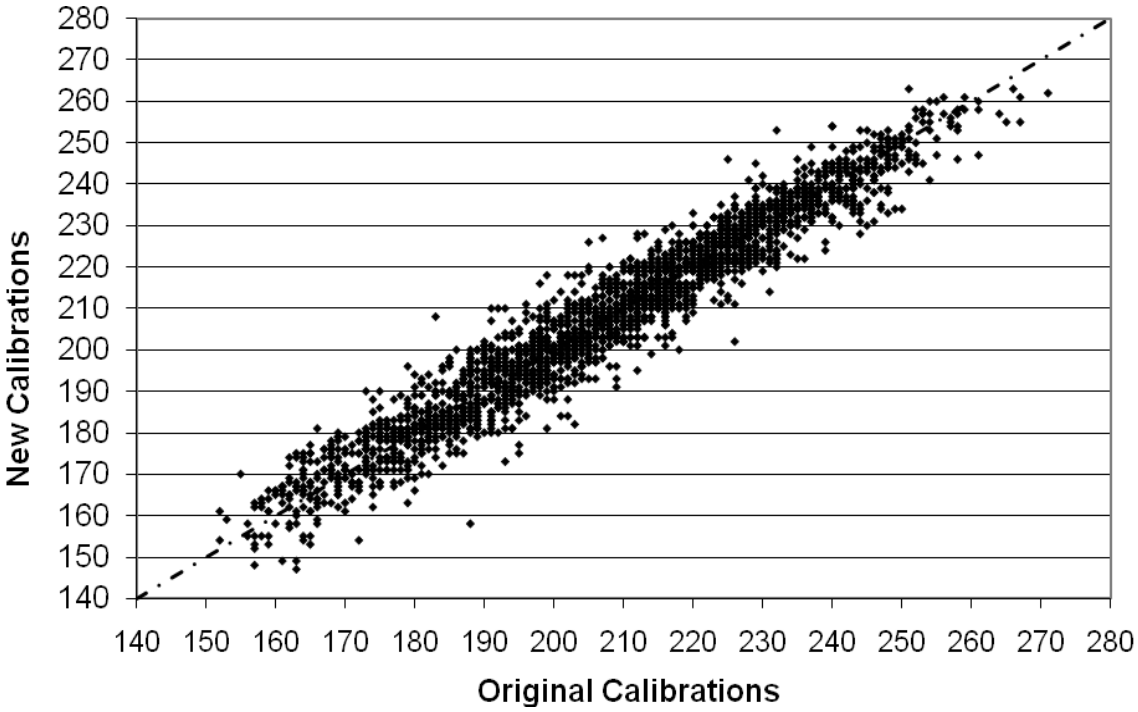


Figure 4.

Relationship of original and new item difficulty estimates on the RIT scale for 1392 items in reading with superimposed identity line (r = .976)

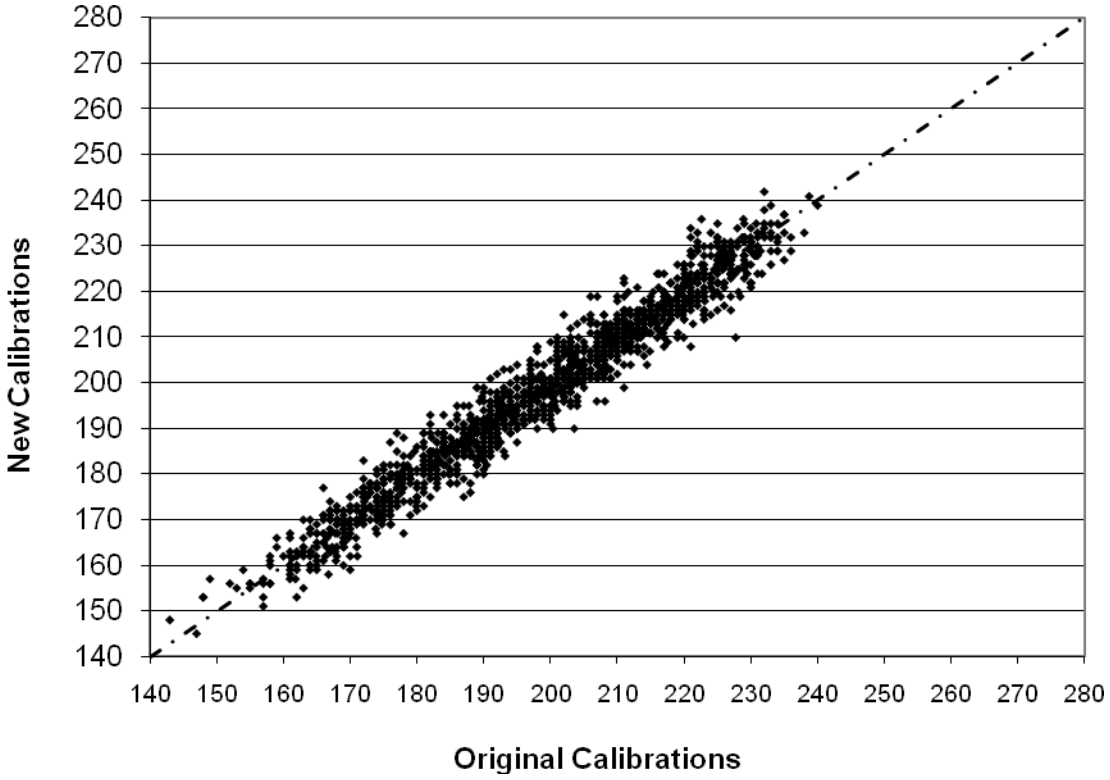


Figure 5.

Differences between original RIT difficulty estimates and new RIT difficulty estimates as a function of initial date of calibration in mathematics

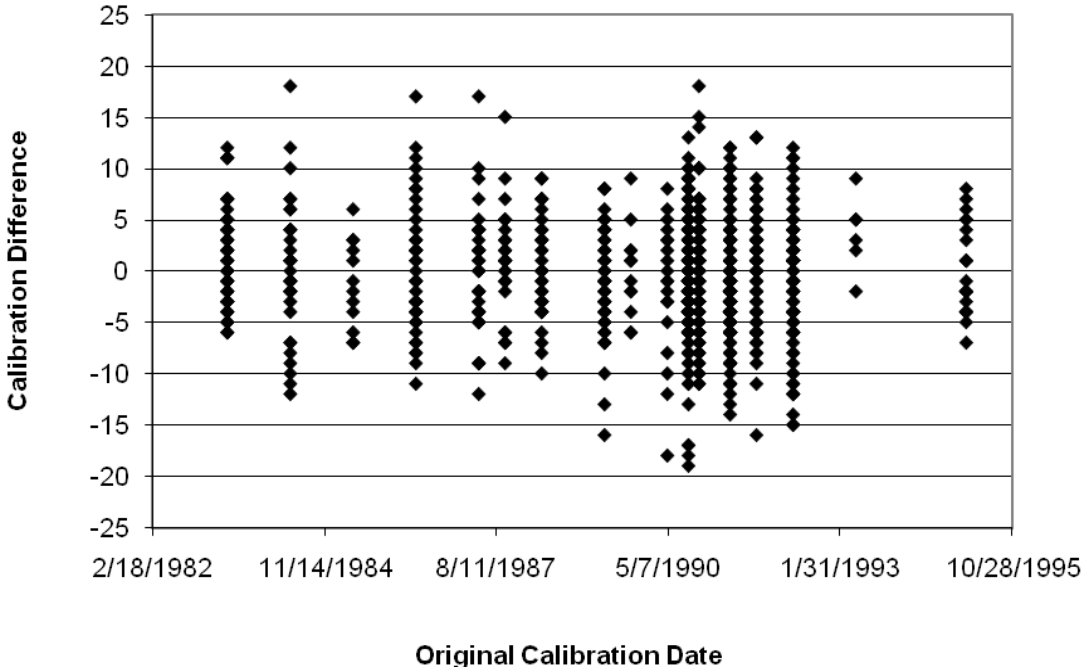


Figure 6.

Differences between original RIT difficulty estimates and new RIT difficulty estimates as a function of initial date of calibration in reading

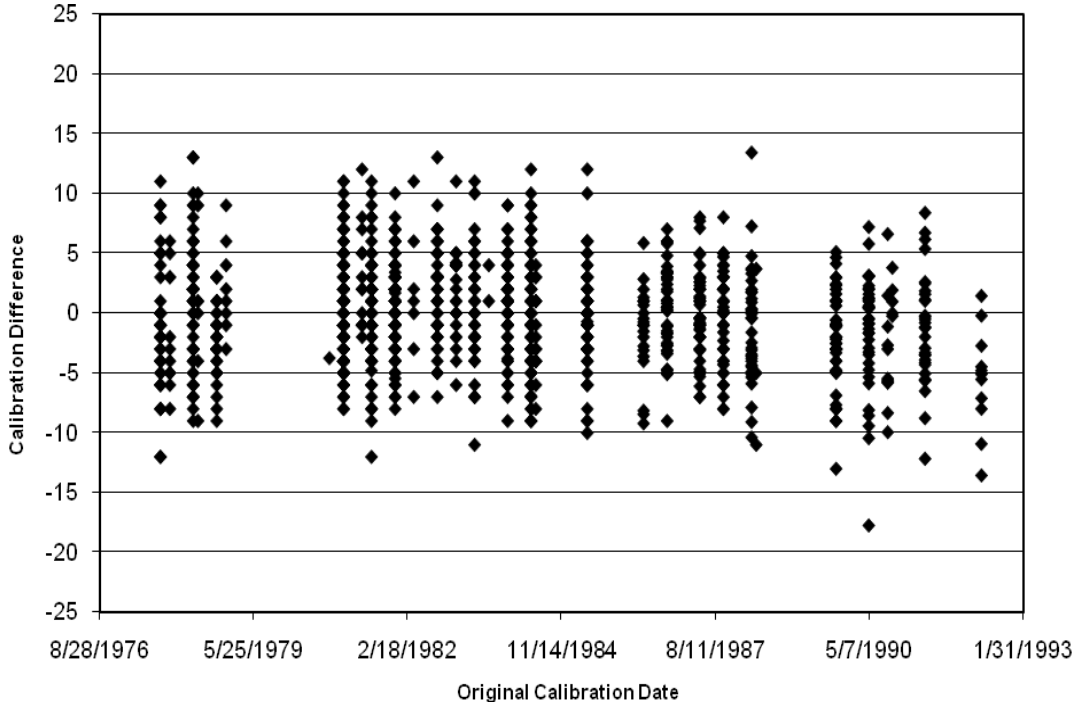


Figure 7.

RIT scores as a function of obtained number correct score calculated using original and new mathematics difficulty estimates

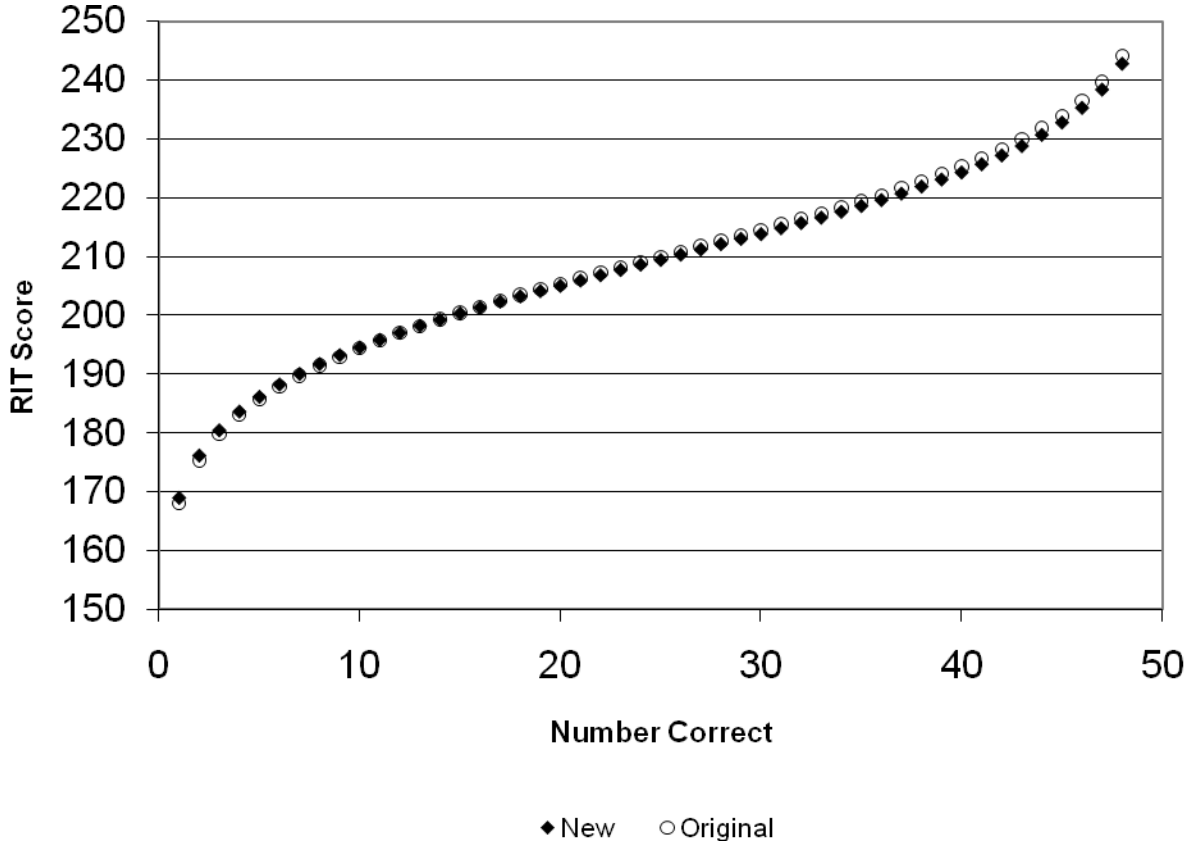
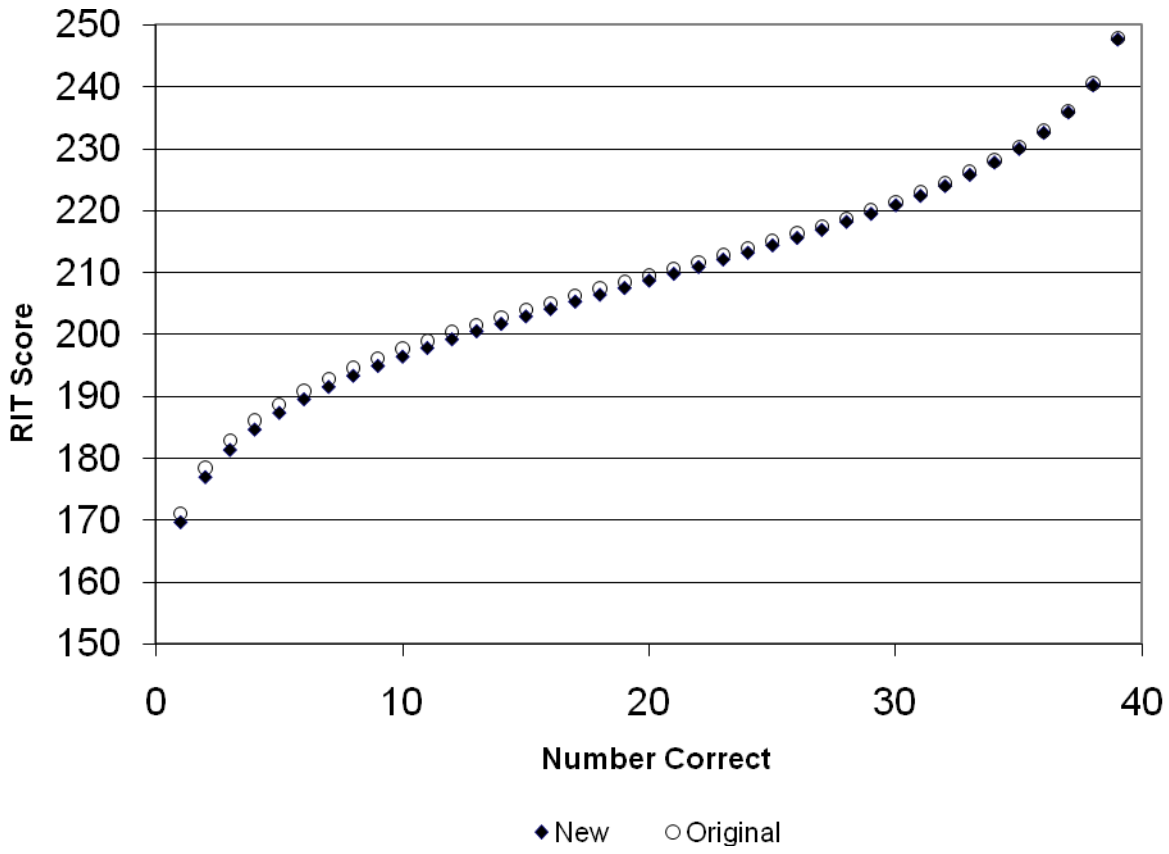


Figure 8.

RIT scores as a function of obtained number correct score calculated using original and new reading difficulty estimates



NWEA™

Northwest Evaluation Association

Partnering to help all kids learn®

(This page left intentionally blank)

Construct Validity and Measurement Invariance of Computerized Adaptive Testing: Application to Measures of Academic Progress (MAP) Using Confirmatory Factor Analysis

Shudong Wang
NWEA

Marty McCall
Smarter Balance Assessment Consortium

Hong Jiao
University of Maryland

Gregg Harris
NWEA

Paper presented at the annual meeting of the American Educational Research Association (AERA). April 12-16, 2012, Vancouver, British Columbia, Canada.

Send correspondence to:
Shudong Wang
Northwest Evaluation Association (NWEA)
121 NW Everett St.
Portland, OR 97206
Shudong.Wang@NWEA.org

Abstract

The purposes of this study are twofold. First, to investigate the construct or factorial structure of a set of Reading and Mathematics computerized adaptive tests (CAT), *Measures of Academic Progress* (MAP), given in different states at different grades and academic terms. The second purpose is to investigate the invariance of test factorial structure across different grades, academic terms and states. Because of the uniqueness of CAT data (different student receive different items), traditional factor analysis based on fixed form data is no longer practically possible at the item level. This study illustrates how to overcome the difficulty of applying factor analysis in CAT data and study results provide evidences for valid interpretation MAP tests scores across grades at different academic terms for different states.

Construct Validity and Measurement Invariance of Computerized Adaptive Testing: Application to Measures of Academic Progress (MAP) Using Confirmatory Factor Analysis

Objectives

The purposes of this study are twofold: first, to investigate the construct or factorial structure of CAT MAP Reading and Mathematics tests at different grades, academic terms, and states; second, to investigate the invariance of test factorial structure across different grades, academic terms and states.

Perspectives

Recently, computerized adaptive testing (CAT) has been seen as a particularly effective method of measuring an individual student's status and growth over time in K-12 assessment (Way, Twing, Camara, Sweeney, Lazer, & Mazzeo, 2010). The major reason is that the CAT has advantages, such as a short test, immediate feedback on student scores, better reliability, and accuracy (Lord, 1977; Kingsbury & Weiss, 1983; Steinberg, & Thissen, 1990) over traditional paper-pencil tests. Its unique advantages in K-12 assessment include cost savings, multiple testing opportunities for formative and interim assessments, and better validity (Way, 2006).

Right now, Oregon, Delaware, and Idaho use CAT in their state assessments, and several other states (Georgia, Hawaii, Maryland, North Carolina, South Dakota, Utah, and Virginia) are in various stages of CAT development. As a matter of fact, one of the two consortia created as part of the Race to the Top initiative, the SMARTER Balanced Assessment Consortium (SBAC) consisting of over half of the states, is committed to a computerized adaptive model because it represents a unique opportunity to create a large-scale assessment system that provides maximally accurate achievement results for each student (Race to the Top Assessment Program, 2010).

Because high stakes decisions about students are based on state test results, these tests should be evaluated using professional testing principles, such as validity and reliability. Validity (and fairness), according to the Standards for Educational and Psychological Testing (AERA, APA, NCME, 1999), is the most important consideration in test development and evaluation.

The MAP Reading and Mathematics tests, like most CATs, use a unidimensional item response model (IRT) model based on the premise that correlations among responses to test questions can be explained by a single underlying trait. Traits like reading and math are

obviously complex, representing many component skills and facts combined in specific ways. The claim of unidimensionality is that these components work together to manifest a coherent whole. Although tests are often structured around goal areas, this is done to provide adequate domain sampling rather than to measure different traits. While individuals may have strengths and weaknesses in goal areas on a unidimensional test, any systematic relationship among goals should be explained by the effect of the unitary latent trait on item responses.

Detecting dimensionality in adaptive assessments is tricky. Because of the uniqueness of CAT tests (different persons respond to different items), conducting factor analysis is more challenging for CAT data than for linear or fixed form data.

First of all, there are no common test forms, so data are very sparse. Observable (manifest) item variables differ across persons for both the overall test event and at the cluster (goal or subtest) level. So, although goal score variables are the same, the context differs; i.e. subtest scores are derived from different sets of items. One possible solution is to conduct confirmatory factor analysis (CFA) on the entire item bank, but the large amount of missing data (typically, the missing rate is above 90% if the ratio of test length to item bank size is 20) makes this unwieldy. For MAP Reading and Mathematics tests, typical missing rates are around 98% because the ratio of test length to item bank counts is around 50. The common imputation methods (Rubin, 1987) may statistically help the missing issue, but are difficult to execute.

The second issue is that the adaptive algorithm operates on the assumption of local independence (LI), thus restricting covariance among items. Items are selected to maximize information at the estimated latent trait level so that for dichotomous items the probability of a correct answer is about .5, responses are randomly distributed, and item covariance is low. Since the goal of factor analysis is to summarize patterns of correlation among observed variables, this restriction may lead to singularly uninformative factor analysis results for CATs. McCall and Hauser (2006) used Yen's (1984) Q statistic to get around the sparse data problem. The Q statistic operates on pairwise relationships between items and looks for covariance unexplained by the observed score. Because item selection is conditioned on the momentary achievement estimate, the range of ability is restricted, thus limiting variance and covariance. Values of the Q statistic were so small compared to those for fixed form tests that they were difficult to interpret.

One way to get around the sparse data problem is to conduct CFA at the item cluster level (goal or sub-content level). Since items in each CAT test event are balanced among goal areas based on content specifications, a reasonable option may be to assume that items within each goal area are content homogeneous across persons and that goal scores may be used as

observable variables. This is the method used here. If multiple traits based on goal areas explain item responses, this might show up as differential factor loading of the goal scores on the overall score. Furthermore, patterns of factor loading might differ with the different goal structures used in different states or among grades within the same state. For illustration purpose only, Figure 1 shows the factor model under the IRT assumption of LI. This is the model that is difficult to test with computerized adaptive test data. Figure 2 shows a factor model that uses testlet, goal score, or item clusters as observable variables. At the testlet level, this model still satisfies the LI assumption, but the LI assumption might or might not hold at the individual item level.

Methods

Data source or Participants

All data used in this study were collected from MAP Reading and Mathematics tests administered from Spring 2009 to Spring 2011 twice during the academic year. The MAP tests were used with grade 3 to grade 9 across 50 states. The data for this study focuses on 10 states (Colorado, Illinois, Indiana, Kansas, Kentucky, Michigan, Minnesota, South Carolina, Washington, and Wisconsin) that have the largest MAP sample sizes among the 50 states. Reading and Mathematics sample sizes for each state are presented in Table 1. Samples were collected for five academic terms: Spring 2009, Fall 2009, Spring 2010, Fall 2010, and Spring 2011. For each academic term, the samples contained results from five grades, with the grade range depending on the academic term. Table 2 and 3 (due to the limited space, can't list 10 state tables) list the frequency and percentages of samples across grades and terms for the Illinois Mathematics test and South Carolina Reading test. For each state, samples were randomly drawn from state records. Approximately 20% students for each state were selected under the constraints that student has to have scores for five academic terms and is in the grade range of 3 to 9 in the first term.

Instruments

The MAP tests of Reading and Mathematics for grades 3 to 9 were used in this study. MAP tests are computerized adaptive assessments that have been published by Northwest Evaluation Association (NWEA) since 2000. The purpose of MAP tests is to provide educators with information to inform teaching and learning in Reading, Mathematics, and Science (NWEA, 2011). For each state, the MAP tests are aligned to specific state content standards by assembling pools of items that address state content standards. Test algorithms survey the pools within goal

or strand areas to assure domain coverage. The marginal reliabilities of tests across 50 states and grades are consistently in the low to mid 0.90's (NWEA, 2011). Because items selected during the CAT test for each student are based on the student's provisional ability, these items have a limited range of difficulty for a given test taker. However, all items administered to each student have to satisfy the content requirements of the test to insure content validity and domain coverage. Table 4 lists test length (fixed length CAT) and numbers of goals (subtests) of both Reading and Mathematics tests. The examples of content specifications for Colorado Reading and Indiana Mathematics are shown in Table 5.

Data Analysis

Using Proc TCALIS in SAS® 9.2 (SAS Institute Inc., 2008), both confirmatory factor analysis (CFA) and multi-group confirmatory factor analysis (MGCFA) were conducted to determine the adequacy of fit of the factor structures of MAP tests and invariance of factor models across grades and academic terms (invariance across terms were not statistically tested). All estimation in this study use the maximum likelihood method.

All MAP tests assume there is only one latent factor (student achievement) that accounts for covariance among observed variables at item or item cluster levels. All MAP tests were scaled based on unidimensional Rasch model (Rasch, 1980) and *RIT* (Rasch unIT) scale that is linearly transformed from logit (*RIT* = *logit* x 10 + 200, NWEA, 2011). Figure 3 present CFA and MGCFA models of MAP tests and the detailed information of the represented models can be found in papers of McArdle (1988) and McDonald (1985).

The one-factor model with goal scores (or subtests) as observed variables and CFA was used to evaluate the adequacy of model to fully account for the relationships among subtests. Once adequacy of model fit was determined, MGCFA was used to test whether the same model holds across different groups. According to Steenkamp and Baumgarther (1998), the invariance of factor loadings is sufficient for construct comparability across groups. In this study, the additional condition of invariance of factor variance was also tested. Three levels of invariance across 5 grades at each of the academic calendars tested are no constraint (NC), equal factor loading (L), and equal factor loadings and factor variances (LV, see Appendix A).

Several well-known goodness-of-fit indexes (GOF) were used to evaluate model fit: (1) absolute indexes that include chi-square χ^2, unadjusted goodness-of-fit indexes (GFI), and standardized root mean square residual (SRMR); (2) incremental indexes that include the

comparative fit index (CFI) and Bentler-Bonett normal fit index (NFI); (3) parsimony index, the root mean square error of approximation (RMSEA). For group comparisons with increased constraints, the χ^2 value provides the basis of comparison with the previously fitted model, although χ^2 is not considered as the best practice because it is sample size dependent. A non-significant difference in χ^2 values between nested models reveals that all equality constraints hold across the groups. Therefore, the measurement model remains invariant across groups as the constraints are increased. A significant χ^2 does not necessarily indicate a departure from invariance when the sample size is large. Hu and Bentler (1999) recommended using combinations of GOF indices to obtain a robust evaluation of data-model fit in structural equation modeling. The cutoff criterion values of good model fit they recommended are CFI, GFI, NFI > 0.95, RMSEA < 0.06, and SRMR < 0.08. It is worth to note that many researchers (March 2007a, Marsh, Hau, & Grayson, 2005) showed that GOF criteria from Hu and Bentler (1999) are too restrictive.

Results

1. *Results of CFA*

Tables 6 and 7 present the summaries of GOF indexes for independent models of Washington MAP Reading and South Carolina MAP Mathematics tests for by grade and term (because of limited space, only partial results are listed for two of 10 states). Although not shown in the table, all factor loadings of models across content, grades, and states are statistically significant. There are mixed results on the statistically significant χ^2 tests (Washington Reading tests are not significant and South Carolina Mathematics are significant) and very similar patterns of χ^2 tests results hold for the rest of the states tests. However, given the large sample sizes across states, it is not surprising to have statistically significant χ^2 tests results for some states. All values of fit indexes (except RMSEAs for Michigan MAP Mathematics tests) satisfy the Hu and Bentler (1999) criteria and show that each model fits data extremely well for different content areas, grades, terms, and states. Overall results suggest that the one-factor (unidimensional) model is the most reasonable model for MAP tests in these 10 states.

2. *Results of MGCFA*

Tables 8 and 9 display the summaries of GOF indexes of the nested models that tested for measurement invariance across grades for Kansas MAP Reading and Michigan Mathematics Tests. In the nested model comparison, the effect of constraints (NC, L, and LV) imposed on

less restricted modes can be evaluated by using the difference between χ^2 (called $\Delta\chi^2$) because it is distributed as χ^2 with the degree of freedom equal to the difference in degrees of freedom between the two models. The null hypothesis of no significant difference in fit is tested by evaluating whether the chi-square difference is significant. If the difference is significant, then the null hypothesis is rejected (Loehlin, 2004). However, the χ^2 test may be misleading because (1) the more complex the model, the more likely a good fit, (2) the larger the sample size, the more likely the rejection of the model and the more likely a Type II error, and (3) the chi-square fit index is also very sensitive to violations of the assumption of multivariate normality. To address these limitations, the difference of other GOF (CFI, GFI, NFI, RMSEA, and SRMR) as adjuncts to the χ^2 statistic can also be used to assess model fit. For the Kansas MAP Reading Tests (see Table 8), χ^2 increases ($\Delta\chi^2$) are significant for testing L invariance at different terms, but not significant for testing LV invariance. The rest of states results show a similar pattern. For Michigan Mathematics Tests, all χ^2 increases are significant for both L and LV invariance. All fit indexes for both Reading and Mathematics tests for different grades and academic years from 10 states satisfied Hu and Bentler's criteria, except RMSEAs and SRMRs for Michigan Mathematics Tests. In summary, the results provide clear support for the metric invariance for all tests except for Michigan Mathematics Tests, and at least, there are configure invariances for all tests.

These results suggest that constructs of MAP tests are well defined, proved to be unidimensional equivalent across grades, and have the same patterns across academic years.

Scientific Significance of the Study

The factor structure of test for a particular grade is directly related to the construct validity interpretation of the test, and validity is one of the most important considerations when evaluating a test. The factor invariance across grades is a fundamental requirement for use in vertical scaling and interpretation of student growth based on the test scores. There are many challenges to providing validity evidence for CAT tests because of its uniqueness compared to fixed form tests. This study using real data provides empirical evidence of construct and invariance construct of MAP scales across grades at different academic calendars for 10 different states. Results show the consistency and reasonableness of interpretation of the MAP RIT scale across grades and academic calendar years for the different states.

Table 1. Sample Sizes for MAP Reading and Mathematics Tests across States

State Name	Reading	Mathematics
Colorado	256310	259600
Illinois	444485	433595
Indiana	262740	247905
Kansas	217730	211070
Kentucky	149785	148725
Michigan	150945	151645
Minnesota	457630	448470
South Carolina	473135	465525
Washington	316980	316925
Wisconsin	351740	351690

Table 2. Frequency and Percentages* of Samples across Grades and Academic Calendars for Illinois MAP Mathematics Test

Grade	Spring 2009	Fall 2009	Spring 2010	Fall 2010	Spring 2011	Total
3	20000					20000
	(4.61)					(4.61)
4	20000	20000	20000			60000
	(4.61)	(4.61)	(4.61)			(13.84)
5	20000	20000	20000	20000	20000	100000
	(4.61)	(4.61)	(4.61)	(4.61)	(4.61)	(23.06)
6	20000	20000	20000	20000	20000	100000
	(4.61)	(4.61)	(4.61)	(4.61)	(4.61)	(23.06)
7	6719	20000	20000	20000	20000	86719
	(1.55)	(4.61)	(4.61)	(4.61)	(4.61)	(20.00)
8		6719	6719	20000	20000	53438
		(1.55)	(1.55)	(4.61)	(4.61)	(12.32
9				6719	6719	13438
				(1.55)	(1.55)	(3.10)
Total	86719	86719	86719	86719	86719	433595
	(20.00	(20.00)	(20.00)	(20.00)	(20.00)	(100.00)

*: Percentage in parentheses

Table 3. Frequency and Percentages* of Samples across Grades and Academic Calendars for South Carolina MAP Reading Test

Grade	Spring 2009	Fall 2009	Spring 2010	Fall 2010	Spring 2011	Total
3	20000					20000
	(4.23)					(4.23)
4	20000	20000	20000			60000
	(4.23)	(4.23)	(4.23)			(12.68)
5	20000	20000	20000	20000	20000	100000
	(4.23)	(4.23)	(4.23)	(4.23)	(4.23)	(21.14)
6	20000	20000	20000	20000	20000	100000
	(4.23)	(4.23)	(4.23)	(4.23)	(4.23)	(21.14)
7	14627	20000	20000	20000	20000	94627
	(3.09)	(4.23)	(4.23)	(4.23)	(4.23)	(20.00)
8		14627	14627	20000	20000	69254
		(3.09)	(3.09)	(4.23)	(4.23)	(14.64)
9				14627	14627	29254
				(3.09)	(3.09)	(6.18)
Total	94627	94627	94627	94627	94627	473135
	(20.00	(20.00)	(20.00)	(20.00)	(20.00)	(100.00)

*: Percentage in parentheses

Table 4. Test Length and Numbers of Goals (subtests) of Reading and Mathematics Tests for Grades 3 to 9 across States

State Name	Reading		Mathematics	
	Test Length	Number of Goal	Test Length	Number of Goal
Colorado	40	4	50	6
Illinois	40	4	50	5
Indiana	40	5	50	7
Kansas	40	5	50	4
Kentucky	40	5	50	5
Michigan	40	4	50	6
Minnesota	40	4	50	4
South Carolina	40	3	50	5
Washington	40	5	50	4
Wisconsin	40	4	50	5

Table 5. Content Specifications of Colorado Reading and Indiana Mathematics for Grades 3 to 9

Colorado Reading Goal	% items per goal	Indiana Mathematics Goal	% items per goal
Reading Strategies, Comprehending Literary Texts	25%	Number Sense	14%
Comprehending Informative and Persuasive Texts	25%	Computation	14%
Word Relationships and Meanings	25%	Algebra and Functions	14%
Total operational items	25%	Geometry	14%
		Measurement	14%
		Statistics, Data Analysis, and Probability	14%
		Problem Solving	14%

Table 6. Summary of Goodness-of- Fit Indexes of Models of Washington MAP Reading Tests for Each Grade at Each Academic Calendar

Academic Calendar	Grade	N	χ^2	df	CFI	GFI	NFI	RMSEA	SRMR
Spring 2009	3	12795	24.94	5	1.00	1.00	1.00	0.02	0.00
	4	13296	7.19	5	1.00	1.00	1.00	0.01	0.00
	5	12957	9.11	5	1.00	1.00	1.00	0.01	0.00
	6	14285	7.98	5	1.00	1.00	1.00	0.01	0.00
	7	10065	3.33	5	1.00	1.00	1.00	0.00	0.00
Fall 2009	4	12795	10.07	5	1.00	1.00	1.00	0.01	0.00
	5	13296	15.13	5	1.00	1.00	1.00	0.01	0.00
	6	12957	12.55	5	1.00	1.00	1.00	0.01	0.00
	7	14285	15.06	5	1.00	1.00	1.00	0.01	0.00
	8	10065	6.52	5	1.00	1.00	1.00	0.00	0.00
Spring 2010	4	12795	17.56	5	1.00	1.00	1.00	0.01	0.00
	5	13296	19.52	5	1.00	1.00	1.00	0.01	0.00
	6	12957	10.78	5	1.00	1.00	1.00	0.00	0.00
	7	14285	5.72	5	1.00	1.00	1.00	0.00	0.00
	8	10065	7.29	5	1.00	1.00	1.00	0.00	0.00
Fall 2010	5	12795	15.39	5	1.00	1.00	1.00	0.01	0.00
	6	13296	5.28	5	1.00	1.00	1.00	0.00	0.00
	7	12957	11.66	5	1.00	1.00	1.00	0.01	0.00
	8	14285	18.77	5	1.00	1.00	1.00	0.01	0.00
	9	10065	6.06	5	1.00	1.00	1.00	0.00	0.00
Spring 2011	5	12795	4.4	5	1.00	1.00	1.00	0.00	0.00
	6	13296	10.63	5	1.00	1.00	1.00	0.01	0.00
	7	12957	15.08	5	1.00	1.00	1.00	0.01	0.00
	8	14285	13.90	5	1.00	1.00	1.00	0.01	0.00
	9	10065	4.52	5	1.00	1.00	1.00	0.00	0.00

Table 7. Summary of Goodness-of- Fit Indexes of Models of South Carolina MAP Mathematics Tests for Each Grade at Each Academic Calendar

Academic Calendar	Grade	N	χ^2	df	CFI	GFI	NFI	RMSEA	SRMR
Spring 2009	3	20000	58.25	5	1.00	0.99	1.00	0.02	0.01
	4	20000	109.70	5	0.99	1.00	1.00	0.03	0.00
	5	20000	152.66	5	1.00	1.00	0.99	0.04	0.00
	6	20000	88.94	5	1.00	0.99	1.00	0.03	0.01
	7	13205	65.05	5	0.99	1.00	1.00	0.03	0.00
Fall 2009	4	20000	70.17	5	0.99	1.00	1.00	0.03	0.00
	5	20000	113.07	5	1.00	0.99	1.00	0.03	0.01
	6	20000	58.18	5	1.00	1.00	0.99	0.02	0.00
	7	20000	88.11	5	1.00	1.00	1.00	0.03	0.00
	8	13205	74.75	5	1.00	1.00	0.99	0.02	0.00
Spring 2010	4	20000	149.51	5	1.00	0.99	1.00	0.04	0.01
	5	20000	180.02	5	1.00	0.99	1.00	0.03	0.01
	6	20000	145.66	5	1.00	1.00	0.99	0.04	0.00
	7	20000	128.38	5	1.00	1.00	1.00	0.04	0.00
	8	13205	78.99	5	1.00	1.00	0.99	0.03	0.00
Fall 2010	5	20000	151.35	5	1.00	0.99	1.00	0.04	0.01
	6	20000	37.83	5	1.00	1.00	1.00	0.02	0.00
	7	20000	70.91	5	1.00	1.00	0.99	0.03	0.00
	8	20000	102.66	5	1.00	0.99	1.00	0.03	0.00
	9	13205	66.81	5	1.00	1.00	1.00	0.03	0.00
Spring 2011	5	20000	140.04	5	1.00	0.99	1.00	0.04	0.01
	6	20000	89.24	5	1.00	1.00	1.00	0.03	0.00
	7	20000	158.49	5	1.00	1.00	1.00	0.04	0.00
	8	20000	201.38	5	1.00	1.00	1.00	0.04	0.01
	9	13205	75.93	5	0.99	1.00	1.00	0.03	0.00

Table 8. Results of Comparisons of Model Invariance of Kansas MAP Reading Tests across Five Grades*

Academic Calendar	Grade/ Group	Model	χ^2	df	$\Delta\chi^2$	CFI	GFI	NFI	RMSEA	SRMR
Spring 2009	G3 to G7	1. NC	67.00	25		1.00	1.00	1.00	0.01	0.00
		2. L	528.76	41	461.76	1.00	1.00	0.99	0.02	0.02
		3. LV	619.04	45	90.28	1.00	0.99	1.00	0.02	0.04
Fall 2009	G4 to G8	1. NC	29.88	25		1.00	1.00	1.00	0.00	0.00
		2. L	225.73	41	195.84	1.00	1.00	1.00	0.02	0.03
		3. LV	258.16	45	32.44	1.00	0.99	1.00	0.03	0.04
Spring 2010	G4 to G8	1. NC	61.99	25		1.00	1.00	1.00	0.01	0.00
		2. L	509.85	41	447.86	1.00	0.99	1.00	0.04	0.03
		3. LV	566.77	45	56.92	0.99	0.99	0.99	0.04	0.05
Fall 2010	G5 to G6	1. NC	27.54	25		1.00	1.00	1.00	0.00	0.00
		2. L	333.64	41	306.10	1.00	0.99	1.00	0.03	0.03
		3. LV	379.73	45	46.09	1.00	1.00	1.00	0.03	0.04
Spring 2011	G5 to G6	1. NC	65.99	25		1.00	1.00	1.00	0.01	0.00
		2. L	353.53	41	287.54	1.00	0.99	1.00	0.03	0.03
		3. LV	487.08	45	133.54	1.00	1.00	1.00	0.03	0.06

* The levels of model constraints restricted to be equal across grades are:
1. NC: No Constraint (Model structure).
2. L: Factor loading .
3. LV: Factor loading + Factor Variance.

Table 9. Results of Comparisons of Model Invariance[1] of Michigan MAP Mathematics Tests across Five Grades[*]

Academic Calendar	Grade/ Group	Model	χ^2	df	$\Delta\chi^2$	CFI	GFI	NFI	RMSEA	SRMR
Spring 2009	G3 to G7	1. NC	1844.03	45		0.98	0.98	0.98	0.08	0.02
		2. L	3410.33	65	1566.30	0.97	0.97	0.96	0.09	0.07
		3. LV	3726.05	69	315.72	0.96	0.96	0.96	0.09	0.12
Fall 2009	G4 to G8	1. NC	1995.99	65		0.98	0.98	0.98	0.08	0.02
		2. L	4008.02	69	2012.02	0.97	0.96	0.97	0.10	0.08
		3. LV	4328.11	45	320.10	0.97	0.96	0.97	0.10	0.13
Spring 2010	G4 to G8	1. NC	2628.19	45		0.98	0.97	0.98	0.09	0.02
		2. L	4699.64	65	2071.45	0.97	0.95	0.97	0.11	0.08
		3. LV	152033.96	69	147334.32	0.97	0.95	0.97	0.11	0.14
Fall 2010	G5 to G6	1. NC	3212.61	45		0.98	0.97	0.98	0.10	0.02
		2. L	3843.60	65	630.99	0.97	0.96	0.97	0.10	0.04
		3. LV	4298.94	69	455.34	0.97	0.95	0.97	0.10	0.14
Spring 2011	G5 to G6	1. NC	2705.84	45		0.98	0.97	0.98	0.10	0.02
		2. L	4254.03	65	1548.19	0.97	0.96	0.97	0.10	0.06
		3. LV	4326.35	69	72.32	0.97	0.96	0.97	0.10	0.08

* The levels of model constraints restricted to be equal across grades are:
1. NC: No Constraint (Model structure only).
2. L: Factor loading .
3. LV: Factor loading + Factor Variance.

References

Bryne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for equivalence of factor covariance and mean structures: The issues of partial measurement invariance. *Psychological Bulletin, 105(3)*, 456-466.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55.

Loehin, J. C. (2004). *Latent variable models: An introduction to factor, path, and structural equation analysis*. Mahwah, NJ: Erlbaum.

Lord, F. M. (1977). A broad-range tailored test of verbal ability. *Applied Psychological Measurement, 1*, 95-100.

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika, 36*, 109-133.

Jöreskog, K. G., & Sörbom, D. (1979). *Advances in factor analysis and structural equation models*. Cambridge, MA: Abt Books.

Kingsbury, G. G., & Weiss, D. J. (1983). A comparison of IRT-based adaptive mastery testing and a sequential mastery testing procedure. In D. J. Weiss (Ed.*), New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 257-238). New York: Academic Press.

Marsh, H.W., Muthén, B., Asparouhov, A., Lüdtke, O., Robitzsch, A., Morin, A.J.S., & Trautwein, U. (2009). Exploratory Structural Equation Modeling, Integrating CFA and EFA: Application to Students' Evaluations of University Teaching. *Structural Equation Modeling, 16*, 439-476.

McCall, M. & Hauser, C. (2006). Item Response Theory and Longitudinal Modeling: The Real World is Less Complicated than We Fear. In R. Lissitz, (Ed.), *Assessing and Modeling Cognitive Development in School.* (pp 143-174). Maple Grove, MN: JAM Press.

McArdle, J. J. (1988). Dynamic but Structural Equation Modeling of Repeated Measures Data. In J. R. Nesselroade & R. B. Cattell (Eds.), *The Handbook of Multivariate Experimental Psychology*. New York: Plenum Press.

McDonald, R. P. (1985). *Factor Analysis and Related Methods*, Hillsdale, NJ: Lawrence Erlbaum Associates.

Northwest Evaluation Association. (2011, January). *Technical manual for Measure of Academic Progress & Measure of Academic Progress for Primary Grades*. Portland, Oregon.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons, New York.

SAS Institute Inc. (2008). *SAS/STAT® 9.2 User's Guide*. Cary, NC: SAS Institute Inc.

Steenkamp, J.B., & Baumgartner, H. (1998). Assessing measurement invariance in crossnational consumer research. *Journal of Consumer Research, 25*: 78-90.

Washington State, on behalf of the SMARTER Balanced Assessment Consortium. (2010). *Race to the Top Assessment Program Application for New Grants*. Retrieved from http://www.k12.wa.us/SMARTER/pubdocs/SBAC_Narrative.pdf.

Way, W. (2006). *Online Testing Research: Information and Guiding Transitions to Computerized Assessments*. A white paper from Pearson Educational Measurement.

Way, W., Twing, J., Camara, W., Sweeney, K., Lazer, S., & Mazzeo, J. (2010). *Some considerations related to the use of adaptive testing for the Common Core Assessments*. Retrieved June 11, 2010, from www.ets.org/s/commonassessments/pdf/ AdaptiveTesting.pdf.

Yen, W. (1984). Effects of Local Item Dependence on the Fit and Equating Performance of the Three-Parameter Logistic Model. Applied Psychological Measurement, 8(2), pp 125-145.

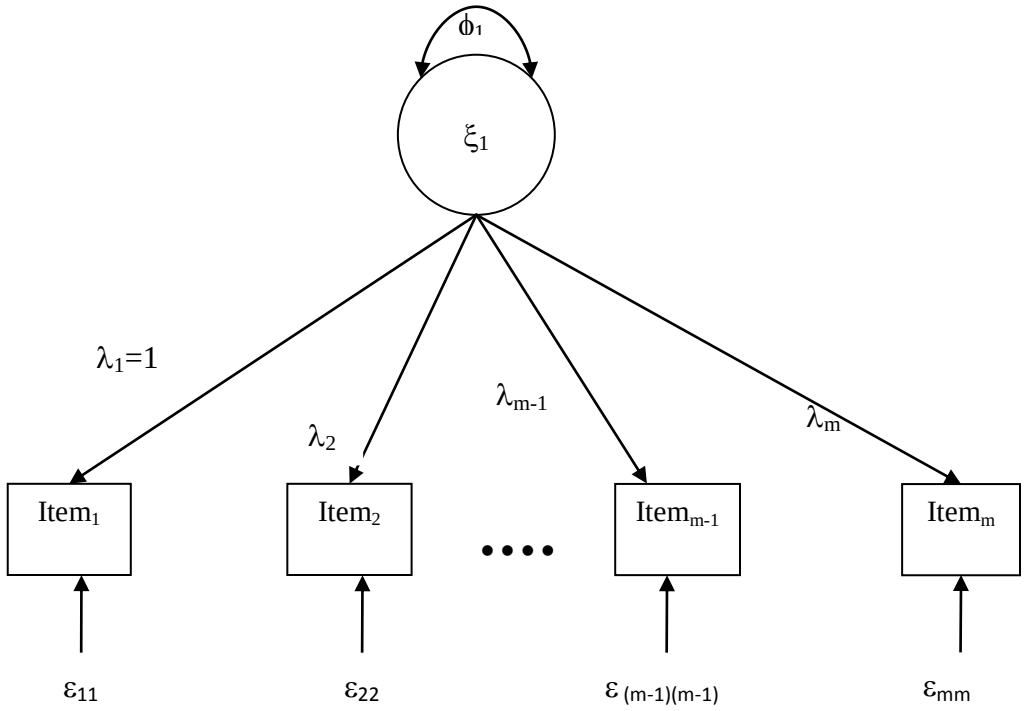


Figure 1. Individual item as observables

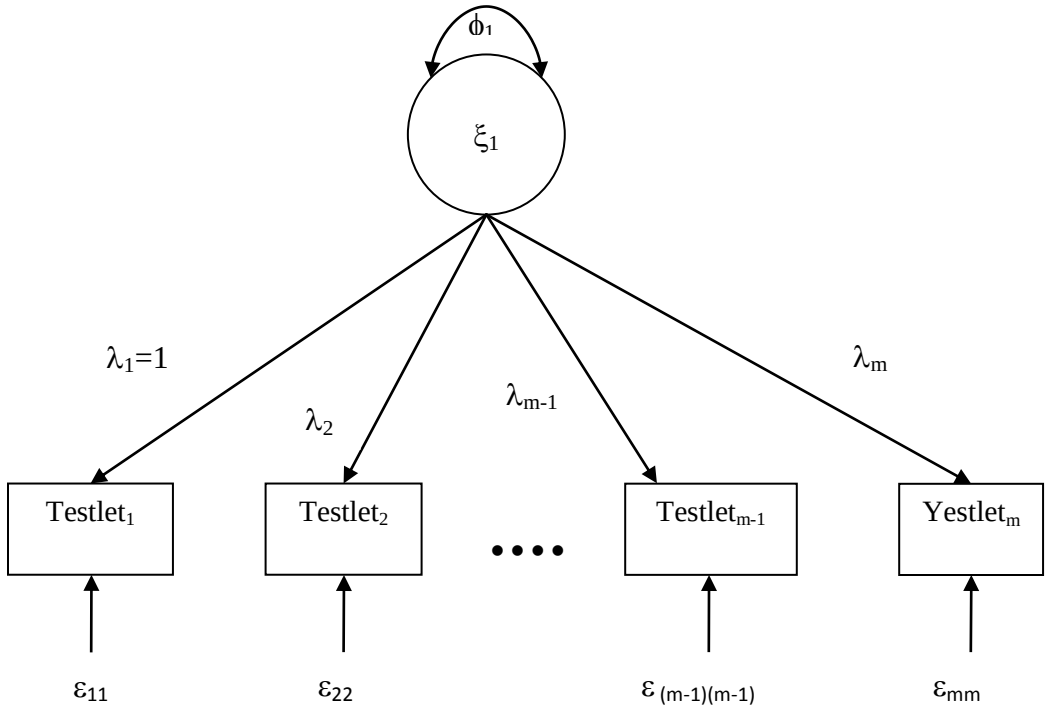
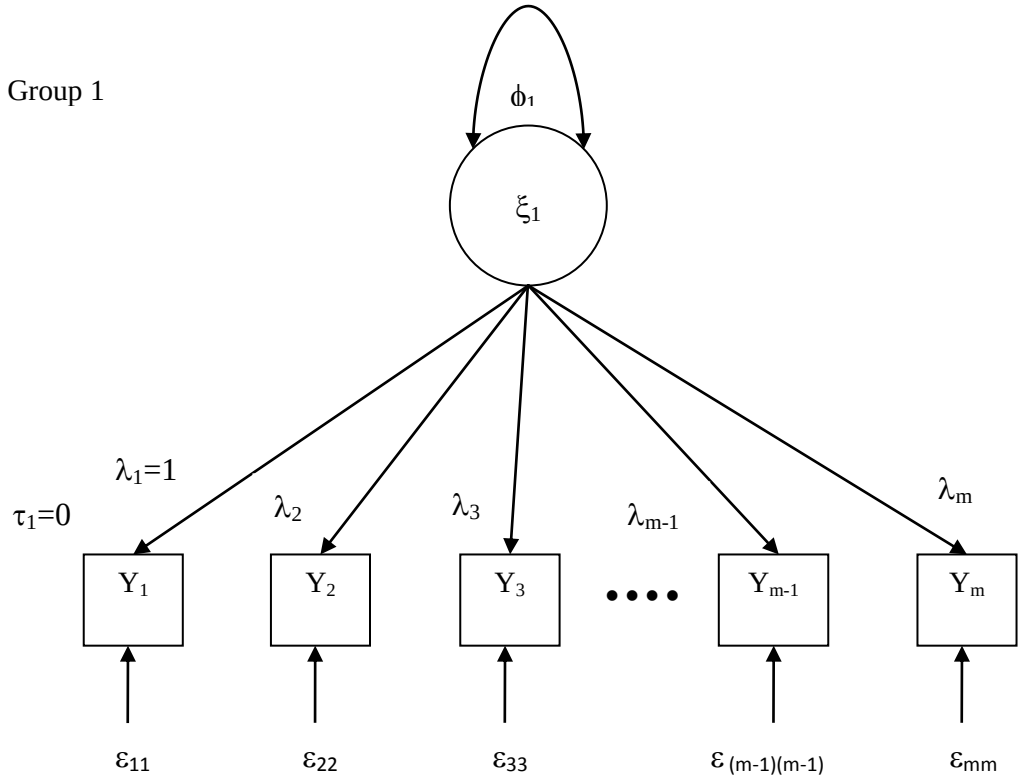


Figure 2. Testlets as observable variables


Group 1
ϕ1
ξ1
λ1=1
τ1=0
λ2
λ3
λm-1
λm
Y1
Y2
Y3
Ym-1
Ym
ε11
ε22
ε33
ε (m-1)(m-1)
εmm

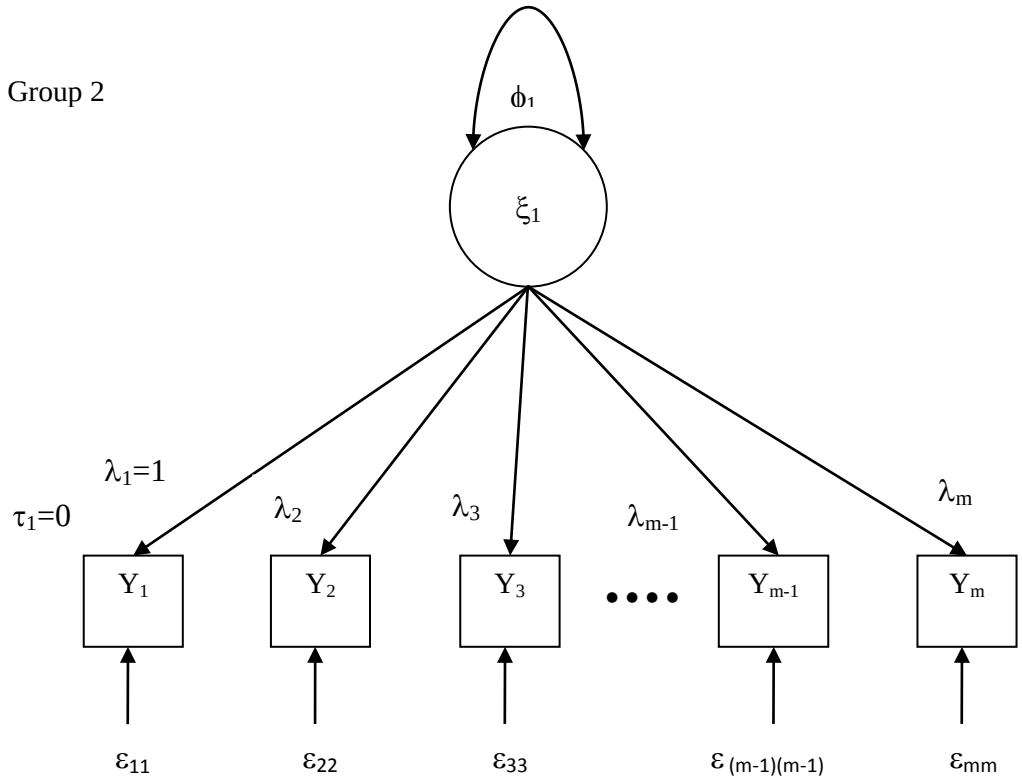

Group 2
ϕ1
ξ1
λ1=1
τ1=0
λ2
λ3
λm-1
λm
Y1
Y2
Y3
Ym-1
Ym
ε11
ε22
ε33
ε (m-1)(m-1)
εmm

Figure 3. Models of MAP Tests across Groups (group indicators omitted for simplicity)

Appendix A.

1. CFA Measurement Model

$$\mathbf{Y} = \boldsymbol{\tau} + \boldsymbol{\Lambda}\boldsymbol{\xi} + \boldsymbol{\varepsilon} \qquad \text{(A.1)}$$

where $\mathbf{Y}$ is the vector of manifest indicator (goal scores in this study), $\boldsymbol{\tau}$ is a vector of measurement intercepts, $\boldsymbol{\Lambda}$ is the matrix of factor loading, and $\boldsymbol{\varepsilon}$ is a vector of residuals. The model-implied covariance matrix is

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Theta} \qquad \text{(A.2)}$$

where $\boldsymbol{\Phi}$ is the latent variable (achievement in this study) covariance matrix and $\boldsymbol{\Theta}$ is the residual covariance matrix. Because we expect the mean achievement will be different across the grades, mean structure is not our concern in this study. In this study, all measurement intercepts were set to zero.

2. CFA Measurement Model in the Multiple Group

$$\mathbf{Y}^{g} = \boldsymbol{\tau}^{\mathbf{g}} + \boldsymbol{\Lambda}^{\mathbf{g}}\boldsymbol{\xi}^{\mathbf{g}} + \boldsymbol{\varepsilon}^{\mathbf{g}} \qquad \text{(A.3)}$$

Where g is group indicator and g = 1,2,...5 in this study. $\mathbf{Y}^{g}$ is the vector of manifest indicator (goal scores in this study), $\boldsymbol{\tau}^{\mathbf{g}}$ is a vector of measurement intercepts, $\boldsymbol{\Lambda}^{\mathbf{g}}$ is the matrix of factor loading, and $\boldsymbol{\varepsilon}^{\mathbf{g}}$ is a vector of residuals. The model-implied covariance matrix is

$$\boldsymbol{\Sigma}^{g} = \boldsymbol{\Lambda}^{\mathbf{g}}\boldsymbol{\Phi}^{\mathbf{g}}\boldsymbol{\Lambda}'^{\mathbf{g}} + \boldsymbol{\Theta}^{\mathbf{g}} \qquad \text{(A.4)}$$

where $\boldsymbol{\Phi}^{\mathbf{g}}$ is the latent variable (achievement in this study) covariance matrix for group g and $\boldsymbol{\Theta}^{\mathbf{g}}$ is the residual covariance matrix for g group. Because we expect the mean achievement will be different across the grades, mean structure is not our concern in this study and all measurement intercepts were set at zero.

According to many researchers (Bryne, Shavelson & Muthén, 1989; Jöreskog ,1971; Marsh, Muthén, Asparouhov, Lüdtke, Robitzsch, Morin &Trautwein, 2009), the invariance of

the parameter matrices implied by equation (A.4) means, the covariance matrices for G groups will only be identical if all of the factor loadings, factor variance and covariances, and residual variance are identical across groups. Although there are total 13 partially nested models (named differently for different researchers) can be tested (Marsh et al., 2009) for model invariance. In this study, three invariance tests conducted are: (1) configure invariance (congeneric invariance) without constraint imposed on parameters; (2) weak factor invariance (tau-equivalent or metric invariance) with constraint of equal factor loading; and (3) invariance of factor loading and factor variance. The invariance tested in this study is summarized as following:

1. No constraint, baseline model (NC)
2. Equal factor loadings (L)

$$H_0: \ \boldsymbol{\Lambda}^1 = \boldsymbol{\Lambda}^2 = \boldsymbol{\Lambda}^3 = \boldsymbol{\Lambda}^4 = \boldsymbol{\Lambda}^5$$

3. Equal factor loadings and factor variance (LV)

$$H_0: \ \boldsymbol{\Lambda}^1 = \boldsymbol{\Lambda}^2 = \boldsymbol{\Lambda}^3 = \boldsymbol{\Lambda}^4 = \boldsymbol{\Lambda}^5$$
$$H_0: \ \boldsymbol{\Phi}^1 = \boldsymbol{\Phi}^2 = \boldsymbol{\Phi}^3 = \boldsymbol{\Phi}^4 = \boldsymbol{\Phi}^5$$

Validation of Longitudinal Achievement Constructs of Vertically Scaled Computerized Adaptive Tests: A Multiple Indicator Latent Growth Modeling Approach

Shudong Wang
NWEA

Hong Jiao
University of Maryland

Liru Zhang
Delaware Department of Education

Send correspondence to:
Shudong Wang
Northwest Evaluation Association (NWEA)
121 NW Everett St.
Portland, OR 97206
Shudong.Wang@NWEA.org

Validation of Longitudinal Achievement Constructs of Vertically Scaled Computerized Adaptive Tests: A Multiple Indicator Latent Growth Modeling Approach

ABSTRACT

It is assumed that an underlying longitudinal achievement construct exists across grades in K-12 achievement tests. This assumption provides a necessary assurance for measurement and interpretation of student growth over time. However, evidence is needed to determine whether the achievement construct remains consistent or shifts over grades or time. The current study is an investigation using a multiple-indicator latent growth modeling (MLGM) approach to examine the longitudinal achievement construct and its invariance for the Measures of Academic Progress® (MAP®) a computerized adaptive tests in reading and mathematics. The results of the analyses from ten states suggest that with repeated measures, the construct of both MAP reading and mathematics are remained consistent at different time points. The findings support the invariance of the achievement construct from different grades or time points and provide empirical validity evidence for measuring student growth.

INTRODUCTION

For decades, achievement tests delivered either in linear or Computerized Adaptive Test (CAT) algorithms have been constructed to provide formative or summative measures about student achievement status at grade level (for example, Amy's third grade mathematics test score in Fall 2005), about students' academic growth over time from longitudinal data (for example, Amy's third grade mathematics test score in Fall 2005 and her seventh grade mathematics score in Fall 2009), or about student achievement across grades from cross-sectional data (for example, Amy's third grade, Johnny's fourth grade, and Tim's fifth grade mathematics scores in Fall 2005) (Hamilton, Stecher & Yuan, 2008; Patz, 2007; Smith and Yen, 2006; Yen, 2007, 2009). In the recent educational reforms, assessing an individual student's growth has been required for high-stakes decisions by state and federal education policy, such as the Race to the Top initiatives (RTTT) (U.S. Department of Education, 2009). These requirements have put tremendous pressures on states and testing companies to develop high-quality and high-utility assessment systems. As RTTT states: "(student) growth may be measured by a variety of approaches, but any approach used must be statistically rigorous and based on student achievement data, and may also include other measures of student learning in order to increase the construct validity and generalizability of the information" (U.S. Department of Education, pp. 37812, 2009). Computerized Adaptive Tests (CAT) are considered more effective than linear tests for measuring individual students' growth over time (Way, Twing, Camara, Sweeney, Lazer, & Mazzeo, 2010). Advantages of CATs over linear tests include shorter tests, immediate reporting of student scores, higher reliability and measurement accuracy (Kingsbury & Weiss, 1983; Lord, 1977; Thissen, & Mislevy, 1990), cost savings, and multiple testing opportunities for formative and interim assessments (Way, 2006).

Measuring individual student growth has two fundamental requirements (Bergman, Eklund, & Magnusson, 1991; Betebenner & Linn, 2010; Dorans & Cohen, 2005; Linn, 1993; Mislevy, 1992;

Williamson, Appelbaum, & Epanchin, 1991). First, there are multiple measures of achievement construct along the growth trajectory. Second, the achievement construct should be invariant from different grade levels or points in time. This means, in order to measure student growth, achievement tests must satisfy two necessary conditions. First, there is a continuous construct that the tests are designed to measure over grades or time, i.e., the tests are scored on a longitudinal construct. Second, the construct measured by multiple tests must be invariant, or constant, across grade levels and time, i.e., the construct does not shift over grade levels or time (Wang & Jiao, 2009).

Whether achievement tests satisfy these two assumptions has been investigated by many researchers from the perspectives of content and vertical scaling (Cizek, 2005; Linn, 2001; Lissitz, 2006; Martineau, Paek, Keene, & Hirsch, 2007; Wang & Jiao, 2009; Wise, 2004). The study by Wang and Jiao (2009) provided empirical evidence to support construct invariance across grades for a vertically scaled norm-referenced test. However, few studies were found that examine the longitudinal achievement construct and construct invariance across grade levels and points in time with CAT.

The invariance of achievement construct is an important validity issue. According to the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999), validity is the most important consideration in test development and evaluation. Validity refers to the degree to which empirical evidence and theoretical rationale support the inferences and actions based on test scores (Messick, 1989) or the degree to which evidence and theory support the interpretations of test scores for the proposed use of tests (AERA, at al., 1999). Interpretations of test results that measure achievement ability or traits are subject to many validity threats. Construct-Under Representation (CUR) and Construct-Irrelevant Variance (CIV) (Messick, 1984) are the major threats to achievement tests. The construct of a test is a theoretical representation of

the underlying traits, concepts, attributes, processes, or structures the test is designed to measure and directly relates to test validity (Cronbach, 1971; Messick, 1989). Five sources of validity evidence specified in the *Standards* include (a) test content, (b) response process, (c) internal structure, (d) relations to other variables, and (e) consequences of testing. CUR and CIV could be identified in the process of test development, administration, and use of test results. This study focuses on validity evidence of internal structure and the invariance of the internal structure of achievement tests across grades or/and over time.

The factorial validity can be a valuable component of validity evidence (Guilford, 1946; Messick, 1995). Although evidence of internal structure of a test is routinely reported in many test technical reports or manuals from state assessment programs and test companies, construct invariance is rarely addressed from a longitudinal perspective when test results are routinely used to determine student growth. Many test publishers have used the First-Order Latent Growth Curve Model (FLGM) (Bollen & Curran, 2006; McArdle, Grimm, Hamagami, Bowles, & Meredith, 2009; Muthén, 1995) in longitudinal studies of student growth based on the assumption that there is a continuous test construct. Unfortunately, little attention has been drawn to examine the assumption in practice. The second-order latent growth model or the Multiple-Indicator Latent Growth Model (MLGM) is rarely used to evaluate the longitudinal construct validity for adaptive test achievement based on CAT. There are two advantages of MLGM over FLGM (Bollen & Curran, 2006; Ferrer, Balluerka, & Widaman, 2008; Muthén & Muthén, 2007; Sayer & Cumsille, 2001). First, MLGM can test, instead of making assumptions, whether the same latent construct is measured at each point in time or grade level. Second, it accounts for important information of the psychometric properties of the indicators. MLGM can be used to evaluate factorial invariance across points in time and determine whether the same latent

construct is measured across time or grade levels so as to assure that changes of test scores quantify achievement growth rather than the shift of the achievement test construct.

FLGM uses total scores as an observable variable or indictor in the longitudinal achievement growth analysis, while MLGM uses observed indicators that could be either at the item level or the cluster of items level (testlet, subtest, goal score, etc.). The choice of the level of an indicator in MLGM or in general factor analysis has a significant effect on evaluating the construct because item cluster involves averaging item scores and using summed scores as observed variables in analyses (Bandalos, 2002; Bandalos & Finney, 2001; Hall, Snell & Foust, 1999; Little, Cunningham, Shahar & Widaman, 2002; Nasser & Wisenbaker, 2003). Some reasons to use clusters of items are to reduce the problems of non-normality, to have fewer free parameters to be estimated compared to the number of observations, and to improve data-model fit. Arguments state that using clusters of items increases the chance of combining items that truly measure multiple dimensions and therefore results in severe bias. Overall, clustering items is a commonly used technique based upon theoretical rationale.

Compared with linear tests, it is almost impossible to use indicators at the item level for adaptive tests because each student receives different items that tailor to his or her ability. Using unique test forms in CAT complicates the evaluation of longitudinal achievement construct. Compared with linear tests, there are two major challenges in CAT of using observable variables for MLGM. First, observable variables are different across persons at the item level. Second, even though the observable variables are the same at the cluster level, the context of the observable variables is different, i.e., the same subtest score may consist of different items for the same content. One possible solution to the first problem is to conduct confirmatory factor analysis at the item level on the entire item bank. However, the drawback is the large amount of missing response data in the dataset. For example, the missing rate of the Measure of Academic Progress (MAP, NWEA, 2011) is around 98% for the reading and mathematics

tests. Besides, the data are very sparse because the ratio of the test length to the size of the item bank is about 1/50 (Wang & Harris, 2011). The commonly used imputation method (Rubin, 1987) may statistically help solve the problem, but it cannot deal with the missing data issue from the content perspective.

The purpose of this study is to investigate the longitudinal achievement constructs of a standardized, large-scale CAT in reading and mathematics across 10 states.

METHOD

Data source and Participants

All data used in this study were collected from the *Measures of Academic Progress* (MAP) assessment system in spring 2009 through spring 2011. The MAP reading and mathematics tests were administered to students in grades 3 through 10 across 50 states in the United States. The data analysis only focused on the 10 states (Colorado, Illinois, Indiana, Kansas, Kentucky, Michigan, Minnesota, South Carolina, Washington, and Wisconsin) that supplied the largest samples among the 50 states. For each state, the data were collected as part of a five-wave panel design (Spring 2009/Grade5, Fall 2009/Grade6, Spring 2010/Grade6, Fall 2010/Grade7, and Spring 2011/Grade7). Each state's sample contains approximately 20% of the students randomly drawn from the corresponding population and under the constraints that students must have five academic calendar records (five-wave) from grade 5.

Tables 1 and 2 list the frequency distributions and percentages of reading and mathematics for sample states (due to limited space) across five-wave by state, gender, and ethnicity.

Insert Tables 1 and 2 about here

Instruments

MAP has been published by Northwest Evaluation Association™ (NWEA™) since 1976, and all MAP tests are computerized and presented adaptively. The purpose of MAP is to provide educators with information to inform teaching and learning in reading, language usage, mathematics, and science (NWEA, 2011). The MAP tests are aligned to the content standards of each state by assembling a customized item pool to measure the specific standards. Unlike state assessment programs used to report proficiency under NCLB, MAP tests allow certain off-grade items to be used for on-grade assessment for the purpose of measuring growth. Test algorithms survey the pool within each content standard goal or strand level to assure the content domain coverage. The marginal reliabilities of test scores are consistently in the range of low to mid 0.90s across grades, tests, and states (NWEA, 2011). MAP items are calibrated with the Rasch model (Rasch, 1961) and all MAP tests are vertically scaled (NWEA, 2011).

In the process of item selection, all items administered to each student must satisfy the content requirements of each test to insure the content validity of the test. Table 3 lists test length (fixed length CAT) and numbers of goals (subtests) for reading and mathematics for the 10 states. A sample of content specifications for reading in Colorado and mathematics in Indiana is shown in Table 4. Due to the uniqueness of CAT test forms, the observed variables used in this study are goal scores (item clusters) under the assumption that the item cluster for each goal area contains homogeneous content across students. All goal scores are scale scores on the same metric across goals and time periods.

Insert Tables 3 and 4 about here

Multiple Indicator Latent Growth Model (MLGM)

MLGM is a multivariate extension of FLGM (Bollen & Curran, 2006; Ferrer, Balluerka, & Widaman, 2008; McDonald, 1985; McArdle, 1988; Muthén, 1991; Tisak & Meredith, 1990). If y_{jti} denote the observed variables (goals) for individual *i*, indicator *j*, and time point *t*, and let η_{ti} denote a latent variable construct, the level-1 model for measurement part is:

$$y_{jti} = \tau_{jt} + \lambda_{jt}\, \eta_{ti} + \varepsilon_{jti} \tag{1}$$

where τ_{jt} is intercept for the *j*th indicator in the *t*th time period, λ_{jt} is the factor loading for the *j*th indicator at the *t*th time point, and ε_{jti} is the random error for the *i*th individual in the *t*th time point and the *j*th indicator. Level-1 models for a latent variable with both linear and quadratic growth are:

Linear growth

$$\eta_{ti} = \eta_{0i} + \eta_{1i}\, \beta_{1t} + \zeta_{ti} \tag{2}$$

Quadratic growth

$$\eta_{ti} = \eta_{0i} + \eta_{1i}\, \beta_{1t} + \eta_{2i}\, \beta_{2t} + \zeta_{ti} \tag{3}$$

where ζ_{ti} is the random normal error for the *i*th individual in the *t*th time point; η_{0i}, η_{1i}, and η_{2i} are the intercept, slope, and quadratic of latent factors, respectively, for individual *i*; and β_{1t} and β_{2t} represent *t*th time point coefficients that determine the shape of the growth curve. Level 2 models are:

Linear growth

$$\eta_{0i} = \alpha_0 + \zeta_{0i} \tag{4}$$

$$\eta_{1i} = \alpha_1 + \zeta_{1i} \tag{5}$$

Quadratic growth

$$\eta_{0i} = \alpha_0 + \zeta_{0i} \tag{6}$$

$$\eta_{1i} = \alpha_1 + \zeta_{1i} \tag{7}$$

$$\eta_{2i} = \alpha_2 + \zeta_{2i} \tag{8}$$

where ζ_{0i} , ζ_{1i} , and ζ_{2i} are normal random errors; α_0, α_1 , and α_2 are latent means of intercept, slope, and quadratic terms for individual *i*, and w_i is weight. All MAP tests have been tested as one latent factor (Wang, McCall, Jiao, & Harris, 2011). Figures 1 and 2 present the MLGM with *linear* and *quadratic* latent growth.

Insert Figure 1 and 2 about here

Measurement Invariance of using MLGM

The measurement invariance (Drasgow, 1987; Ferrer, Balluerka, & Widaman, 2008; Meredith, 1993) evaluated in this study is the invariance across time points of testing. Although values of manifest variables are different across time in longitudinal studies, they should be on the same measurement scale to derive an equal definition of a latent construct across time. Widaman and Reise (1997) classified two types of factorial invariance as non-metric (configural) and metric. Configural invariance (CI) refers to the same indicators of the latent construct. The metric factorial invariance has three hierarchical levels, which are categorized as Weak Invariance (WI), where the factor loading of each indicator is invariant over time; Strong Invariance (SI), where the factor loading and intercept of each indicator are invariant over time; and Strict Factorial Invariance (SFI), where the factor loading, intercept, and unique variance of each indicator are invariant over time. Sayer and Cumsille (2001) showed that the SFI is unlikely to hold because heterogeneous variance across time is often observed. In this study, only CI, WI, and SI are analyzed. The invariance tested is summarized in three conditions: the configural invariance (CI), the weak invariance (WI) that can be expressed as Equation 9, and the strong invariance (SI) that can be expressed in Equations 10 and 11

$$H_0: \lambda_{j1} = \lambda_{j2} = \ldots \lambda_{jT} = \lambda_j \quad (9)$$

$$H_0: \lambda_{j1} = \lambda_{j2} = \ldots \lambda_{jT} = \lambda_j \quad (10)$$

$$H_0: \alpha_{j1} = \alpha_{j2} = \ldots \alpha_{jT} = \alpha_j \quad (11)$$

where $\lambda_1 = 1$, $\alpha_0 = 0$, and variances of ε_{jti} and ζ_{ti} may vary over time. And for structural differences, the mean of η_{ti} and variance of η_{ti} vary over time.

Several well-known goodness-of-fit (GOF) indexes were used to evaluate the model fit. They are (1) absolute indexes that include chi-square χ^2 and standardized root mean square residual (SRMR); (2) incremental indexes that include the comparative fit index (CFI) and Tucker-Lewis Index (TLI); and (3) parsimony index, the root mean square error of approximation (RMSEA). For nested models that include a different shape of growth (e.g., linear versus quadratic), both Akaike's Information Criterion (AIC; Akaike, 1987) and Bayesian Information Criterion (BIC) (Schwartz, 1978) are obtained for each model tested. According to Raftery (1995), the values of BIC difference (BIC of quadratic model – BIC of linear model) ranging from 0 to 2 are interpreted as weak evidence for quadratic model, values of 2 to 6 are interpreted as moderate, values of 6 to 10 are interpreted as strong, and values > 10 are interpreted as very strong. For model comparisons with increased constraints, the χ^2 value also provides the basis of comparison with the previously fitted model in addition to AIC and BIC. χ^2 is not considered as the best one in practice because it is sample-size dependent. A non-significant difference in χ^2 values between nested models reveals that all equality constraints hold across time. Therefore, the measurement model remains invariant across groups as the constraints are increased. A significant χ^2 value does not necessarily indicate a departure from invariance when the sample size is large. Hu and Bentler (1999) recommended using combinations of GOF indices to obtain a robust evaluation for model-data fit in structural equation modeling. The cut-off criterion values of a good model-fit that they recommended are CFI > 0.95, TLI > 0.95, RMSEA < 0.06, and SRMR < 0.08. It is worth noting that many researchers (Hau, & Grayson, 2005; March 2007a, Marsh, 2007; Sayer & Cumsille, 2001) consider the GOF criteria from Hu and Bentler too restrictive. All analyses are conducted in *Mplus* 5.1 (Muthén & Muthén, 2006).

RESULTS

Results of MLGM

Figures 3 and 4 illustrate a sample of observed and estimated individual quadratic growth based on MGLM across five-wave academic terms. Tables 5 and 6 display the summaries of GOF indexes of MLGM data fit for linear and quadratic growth in reading and mathematics across states. All values of the fit indexes satisfy the Hu and Bentler (1999) criteria in both content areas and show that each model fits the data extremely well across states with one exception in Indiana for mathematics.

Insert Figures 3 and 4 about here

Insert Tables 5 and 6 about here

The overall results suggest that both linear and quadratic MLGMs are reasonably good models for MAP tests in 10 states. For AIC, the lower value (positive or negative) indicates a better fit than the higher value. The results show that the quadratic model fits data better than the linear model in the nested modeling comparison. For BIC, all differences between quadratic and linear models are greater than 10 in both reading and mathematics, which indicate that the quadratic model fits the data better than the linear model. The statistically significant χ^2 difference between the quadratic and linear models provides additional evidence to support the conclusion that the quadratic model is a better fit for the data than the linear model. It is also important to note that both the linear and quadratic models show longitudinal achievement construct underlying the achievement measures equally well in growth.

Results of Invariance of MLGM

Tables 7 and 8 present the summaries of GOF indexes with nested linear MLGM that was used for measurement invariance across five waves in reading and mathematics. Nearly all fit indexes satisfied Hu and Bentler's criteria in both reading and mathematics tests across states except Indiana. Some SRMRs seemed to be slightly above Hu and Bentler's criteria. In evaluating measurement invariance, the simple model is a restricted model and the complex model is an unrestricted model. The effect of constraints imposed on the less restricted model can be evaluated by using the difference of χ^2 ($\Delta\chi^2$) for nested model comparisons because the degree of freedom is equal to the difference in the degrees of freedom of two models. Results indicate that all χ^2 increases ($\Delta\chi^2$) are statistically significant for evaluating the differences of invariance between unrestricted and restricted models. As χ^2 becomes statistically significant, a more complex model should be chosen. However, the limitations of the χ^2 test are the sample size dependency (Cheung & Rensvold, 2002), and the difference of other GOF indexes (such as CFI) as adjuncts to the χ^2 statistic, which can also be used to assess model fit. According to Cheung and Rensvold (2002), if the difference of CFI (ΔCFI) is less than 0.01 between the two models, the simple model is not worse than the complex model. The value of all ΔCFIs less than 0.01 in both tests indicates that constrained parameters are invariant across time.

Insert Table 7 and 8 about here

In summary, the results of analyses in this study provide clear support to the CI, WI, and SI for all tests except the Indiana mathematics test. These results suggest that longitudinal constructs of MAP tests are well defined for measuring student achievement growth.

DISCUSSION

As the factor structure of a test is directly related to the construct validity interpretation of the test at a particular point in time, the longitudinal factor structure at different points in time is crucial for the longitudinal construct validity interpretation to measuring student growth. The achievement construct of a test at a particular time, grade level, or semester calendar is well studied and reported in practice for given purposes and related interpretation of test scores. Although many standardized achievement tests in large-scale assessments report test scores on a vertical scale for student growth and group achievement trends, few studies reported the longitudinal achievement construct. Many researchers are interested in whether the longitudinal achievement construct remains the same over time or shifts from time to time from content standard and vertical scaling perspectives. A few studies have focused on validation of longitudinal achievement construct using the MLGM approach, especially for studies based on CAT longitudinal data.

The present study tested the hypothesis of factorial invariance of MAP reading and mathematics tests over time. The evidence collected in the study shows that with repeated measures, the construct of both reading and mathematics remained consistent at different points in time, which supports the internal structure of MAP design for intended purposes. The evidence also suggest that there are not only configural and weak invariance, but also strong invariance of the longitudinal construct in MAP reading and mathematics tests across different states (except the Indiana mathematics test), which supports valid interpretations of student growth.

In summary, this study underscores the importance of empirical evidence in the validation of longitudinal achievement constructs to support the interpretation of student growth. In particular, the study explored the feasibility of assessing the internal structure of MAP tests using CAT data. The

results support consistent and reasonable interpretations of the MAP reading and mathematics tests across academic calendar years used by different states. This study carries the validation process beyond a traditional construct validation process in which validation evidence is usually collected at one point in time only, but used to support the longitudinal achievement construct for student growth. It is indeed important to investigate the longitudinal achievement construct to ensure that the same construct is measured over time for a valid interpretation of student achievement growth. We strongly recommend that achievement test publishers and users continue investigating the longitudinal achievement construct and construct invariance over time in the near future to support valid interpretations of student academic growth.

TABLE 1

Samples of MAP Reading Tests for 10 States

State	Gender		Ethnicity* 1	2	3	4	5	6	7	8	Total
Colorado	F	N	78	51	119	729	4	1351	57	93	2482
		%	1.53	1.00	2.34	14.34	0.08	26.58	1.12	1.83	48.84
	M	N	59	50	113	708	6	1536	52	76	2600
		%	1.16	0.98	2.22	13.93	0.12	30.22	1.02	1.50	51.16
Illinois	F	N	58	754	698	2299	13	5486	220	342	9870
		%	0.29	3.77	3.49	11.50	0.07	27.43	1.10	1.71	49.35
	M	N	53	692	725	2320	15	5733	259	333	10130
		%	0.27	3.46	3.63	11.60	0.08	28.67	1.30	1.67	50.65
Indiana	F	N	20	77	365	254		4247	162	124	5249
		%	0.19	0.74	3.50	2.43		40.71	1.55	1.19	50.31
	M	N	9	60	295	249	1	4285	156	129	5184
		%	0.09	0.58	2.83	2.39	0.01	41.07	1.50	1.24	49.69
Kansas	F	N	189	130	226	355	8	3236	103	105	4352
		%	2.13	1.46	2.54	3.99	0.09	36.38	1.16	1.18	48.93
	M	N	204	138	256	404	3	3329	105	103	4542
		%	2.29	1.55	2.88	4.54	0.03	37.43	1.18	1.16	51.07
Kentucky	F	N	4	41	362	95	1	2491	26	264	3284
		%	0.06	0.63	5.55	1.46	0.02	38.21	0.40	4.05	50.37
	M	N	5	21	337	98	6	2446	37	286	3236
		%	0.08	0.32	5.17	1.50	0.09	37.52	0.57	4.39	49.64
Michigan	F	N	25	100	642	139	6	2295	9	159	3375
		%	0.37	1.46	9.40	2.03	0.09	33.59	0.13	2.33	49.39
	M	N	30	80	582	109	7	2454	3	193	3458
		%	0.44	1.17	8.52	1.60	0.10	35.91	0.04	2.82	50.61
Minnesota	F	N	203	439	607	551	1	7883	2	262	9948
		%	1.02	2.20	3.04	2.76	0.01	39.43	0.01	1.31	49.75
	M	N	206	448	616	529	2	8015	6	224	10046
		%	1.03	2.24	3.08	2.65	0.01	40.09	0.03	1.12	50.25
South Carolina	F	N	24	118	3542	518	12	5514	226	2	9956
		%	0.12	0.59	17.71	2.59	0.06	27.57	1.13	0.01	49.78
	M	N	22	101	3561	545	9	5585	216	5	10044
		%	0.11	0.51	17.81	2.73	0.05	27.93	1.08	0.03	50.22
Washington	F	N	156	228	169	1765	32	3558	169	214	6291
		%	1.20	1.76	1.30	13.62	0.25	27.46	1.30	1.65	48.55
	M	N	153	270	177	1905	51	3739	181	190	6666
		%	1.18	2.08	1.37	14.70	0.39	28.86	1.40	1.47	51.45
Wisconsin	F	N	126	240	462	525	3	5755	8	226	7345
		%	0.84	1.61	3.10	3.52	0.02	38.56	0.05	1.51	49.21
	M	N	99	237	428	527	1	6044	6	239	7581
		%	0.66	1.59	2.87	3.53	0.01	40.49	0.04	1.60	50.79

*1. Native American/Alaskan Native, 2.Asian, 3. African American, 4. Hispanic, 5. Native Hawaiian or Other Pacific Islander, 6. White, 7. Multi-ethnic, 8. Not Specified or Other.

TABLE 2

Samples of MAP Mathematics Tests across Five-wave from 10 States

State	Gender		Ethnicity* 1	2	3	4	5	6	7	8	Total
Colorado	F	N	78	52	123	805	4	1396	60	92	2610
		%	1.44	0.96	2.27	14.86	0.07	25.77	1.11	1.70	48.18
	M	N	64	53	131	830	6	1594	51	78	2807
		%	1.18	0.98	2.42	15.32	0.11	29.43	0.94	1.44	51.82
Illinois	F	N	52	739	744	2186	14	5543	230	360	9868
		%	0.26	3.70	3.72	10.93	0.07	27.72	1.15	1.80	49.34
	M	N	63	680	745	2152	11	5866	260	355	10132
		%	0.32	3.40	3.73	10.76	0.06	29.33	1.30	1.78	50.66
Indiana	F	N	22	76	370	222	0.00	4185	163	128	5166
		%	0.21	0.74	3.61	2.17	0.00	40.84	1.59	1.25	50.41
	M	N	9	61	298	216	1	4214	152	130	5081
		%	0.09	0.60	2.91	2.11	0.01	41.12	1.48	1.27	49.59
Kansas	F	N	180	129	238	356	8	3285	102	108	4406
		%	2.01	1.44	2.65	3.97	0.09	36.64	1.14	1.20	49.14
	M	N	203	137	271	395	4	3337	107	106	4560
		%	2.26	1.53	3.02	4.41	0.04	37.22	1.19	1.18	50.86
Kentucky	F	N	4	42	381	101	1	2486	27	278	3320
		%	0.06	0.64	5.77	1.53	0.02	37.66	0.41	4.21	50.287
	M	N	5	22	340	102	6	2478	36	293	3282
		%	0.08	0.33	5.15	1.54	0.09	37.53	0.55	4.44	49.712
Michigan	F	N	22	86	627	172	5	2301	9	136	3358
		%	0.32	1.26	9.18	2.52	0.07	33.70	0.13	1.99	49.19
	M	N	24	74	563	137	6	2479	3	183	3469
		%	0.35	1.08	8.25	2.01	0.09	36.31	0.04	2.68	50.81
Minnesota	F	N	217	432	628	518	1	7897	2	254	9949
		%	1.09	2.16	3.14	2.59	0.01	39.49	0.01	1.27	49.75
	M	N	225	457	600	523	2	8026	3	215	10051
		%	1.13	2.29	3.00	2.62	0.01	40.13	0.02	1.08	50.26
South Carolina	F	N	22	110	3622	521	11	5477	227	2	9992
		%	0.11	0.55	18.11	2.605	0.055	27.385	1.135	0.01	49.96
	M	N	24	102	3532	540	6	5591	205	8	10008
		%	0.12	0.51	17.66	2.7	0.03	27.95	1.025	0.04	50.04
Washington	F	N	167	264	189	1881	31	3985	172	265	6954
		%	1.17	1.85	1.32	13.17	0.22	27.91	1.20	1.86	48.70
	M	N	178	276	188	2004	54	4205	186	233	7324
		%	1.25	1.93	1.32	14.04	0.38	29.45	1.30	1.63	51.30
Wisconsin	F	N	131	241	481	549	4	5933	19	214	7572
		%	0.85	1.57	3.13	3.58	0.03	38.64	0.12	1.39	49.31
	M	N	102	236	449	539	4	6212	14	228	7784
		%	0.66	1.54	2.92	3.51	0.03	40.45	0.09	1.48	50.69

*1. Native American/Alaskan Native, 2.Asian, 3. African American, 4. Hispanic, 5. Native Hawaiian or Other Pacific Islander, 6. White, 7. Multi-ethnic, 8. Not Specified or Other.

TABLE 3

Test Length and Numbers of Goals (subtests) of Reading and Mathematics Tests for Grades 5 to 7 across States

State	Reading		Mathematics	
	Test Length	Number of Goal	Test Length	Number of Goal
Colorado	40	4	50	6
Illinois	40	4	50	5
Indiana	40	5	50	7
Kansas	40	5	50	4
Kentucky	40	5	50	5
Michigan	40	4	50	6
Minnesota	40	4	50	4
South Carolina	40	3	50	5
Washington	40	5	50	4
Wisconsin	40	4	50	5

TABLE 4

Content Specifications of Colorado Reading and Indiana Mathematics for Grades 5 to 7 across States

Colorado Reading Goal	% items per goal	Indiana Mathematics Goal	% items per goal
Reading Strategies, Comprehending Literary Texts	25%	Number Sense	14%
Comprehending Informative and Persuasive Texts	25%	Computation	14%
Word Relationships and Meanings	25%	Algebra and Functions	14%
Total operational items	25%	Geometry	14%
		Measurement	14%
		Statistics, Data Analysis, and Probability	14%
		Problem Solving	14%

TABLE 5

Goodness-of- Fit Indexes of MLGM Models of MAP Reading Tests for Different States

State	Model	N	χ^2	df	$\Delta\chi^2$	CFI	TLI	RMSEA	SRMR	AIC	BIC
Colorado	Linear	5082	619.234	170		0.996	0.996	0.023	0.044	741066.826	741458.834
	Quadratic		537.906	166	81.328	0.997	0.996	0.021	0.043	740993.498	741411.640
Illinois	Linear	20000	3039.390	170		0.993	0.992	0.029	0.050	2890494.433	2890968.642
	Quadratic		2783.299	166	256.091	0.994	0.993	0.028	0.049	2890246.342	2890752.165
Indiana	Linear	10433	2056.374	275		0.992	0.991	0.025	0.040	1928420.139	1928964.094
	Quadratic		1979.157	271	77.217	0.992	0.992	0.025	0.041	1928350.922	1928923.888
Kansas	Linear	8894	2110.578	275		0.991	0.990	0.027	0.043	1634280.583	1634812.568
	Quadratic		2090.995	271	19.583	0.991	0.990	0.027	0.043	1634269.000	1634829.357
Kentucky	Linear	6520	1179.054	275		0.994	0.994	0.022	0.029	1206322.986	1206831.683
	Quadratic		1082.308	271	96.746	0.995	0.994	0.021	0.022	1206234.240	1206770.068
Michigan	Linear	6833	1299.092	170		0.992	0.991	0.031	0.052	989251.073	989660.844
	Quadratic		1106.289	166	192.803	0.993	0.992	0.029	0.050	989066.269	989503.359
Minnesota	Linear	19994	3713.725	170		0.991	0.990	0.032	0.043	2878450.276	2878924.468
	Quadratic		3466.721	166	247.004	0.992	0.991	0.032	0.046	2878211.273	2878717.077
South Carolina	Linear	20000	4609.603	90		0.988	0.986	0.050	0.076	2123841.525	2124197.182
	Quadratic		3456.152	86	1153.451	0.991	0.989	0.044	0.038	2122696.074	2123083.345
Washington	Linear	12957	2033.200	275		0.995	0.994	0.022	0.026	2385763.775	2386323.979
	Quadratic		1923.339	271	109.861	0.995	0.995	0.022	0.025	2385661.914	2386251.996
Wisconsin	Linear	14926	2718.167	170		0.992	0.991	0.032	0.063	2148535.797	2148992.449
	Quadratic		2490.324	166	227.843	0.992	0.991	0.031	0.062	2148315.954	2148803.050

TABLE 6

Goodness-of- Fit Indexes of MLGM Models of MAP Mathematics Tests for Different States

State	Model	N	χ^2	df	$\Delta\chi^2$	CFI	TLI	RMSEA	SRMR	AIC	BIC
Colorado	Linear	5417	3200.482	405		0.986	0.985	0.036	0.033	1202562.013	1203155.770
	Quadratic		2983.488	401	216.994	0.987	0.986	0.034	0.033	1202353.011	1202973.156
Illinois	Linear	20000	11337.522	275		0.982	0.981	0.045	0.029	3618796.729	3619389.491
	Quadratic		10322.301	271	1015.221	0.984	0.982	0.043	0.027	3617789.507	3618413.883
Indiana	Linear	10247	148965.438	560		0.580	0.553	0.161	3.559	2790203.181	2790962.829
	Quadratic		71963.157	556	77002.281	0.798	0.784	0.112	0.995	2713208.900	2713997.487
Kansas	Linear	8966	4764.568	170		0.979	0.977	0.055	0.048	1264358.923	1264784.995
	Quadratic		3921.295	166	843.273	0.983	0.980	0.050	0.045	1263523.650	1263978.127
Kentucky	Linear	6602	3465.561	275		0.983	0.982	0.042	0.020	1211001.089	1211510.724
	Quadratic		2988.207	271	477.354	0.986	0.984	0.039	0.018	1210531.735	1211068.550
Michigan	Linear	6827	4184.200	275		0.981	0.979	0.046	0.024	1245720.842	1246232.990
	Quadratic		3822.681	271	361.519	0.983	0.981	0.044	0.024	1245367.323	1245906.785
Minnesota	Linear	20000	15870.194	170		0.970	0.966	0.068	0.093	2857014.335	2857488.544
	Quadratic		13337.092	166	2533.102	0.975	0.971	0.063	0.082	2854489.233	2854995.056
South Carolina	Linear	20000	13233.352	275		0.980	0.978	0.049	0.027	3649802.550	3650395.312
	Quadratic		11124.677	271	2108.675	0.983	0.981	0.045	0.021	3647701.876	3648326.251
Washington	Linear	14278	8230.010	170		0.978	0.976	0.058	0.046	2049743.767	2050197.755
	Quadratic		7615.228	166	614.782	0.980	0.977	0.056	0.042	2049136.985	2049621.239
Wisconsin	Linear	15356	9389.812	275		0.980	0.978	0.046	0.024	2780059.981	2780632.926
	Quadratic		8310.063	271	1079.749	0.982	0.980	0.044	0.022	2778988.233	2779591.735

TABLE 7

Goodness-of- Fit Indexes of Invariances of Linear MLGM Models of MAP Reading Tests

State	Model*	N	χ^2	df	$\Delta\chi^2$	CFI	TLI	RMSEA	SRMR	AIC	BIC
Colorado	CI	5082	619.234	170		0.996	0.996	0.023	0.044	741066.826	741458.834
	WI		764.490	182	145.256	0.995	0.995	0.025	0.092	741188.082	741501.688
	SI		1264.094	194	499.604	0.991	0.991	0.033	0.096	741663.686	741898.891
Illinois	CI	20000	3039.390	170		0.993	0.992	0.029	0.050	2890494.433	2890968.642
	WI		3220.062	182	180.672	0.993	0.992	0.029	0.073	2890651.105	2891030.472
	SI		6975.913	194	3755.851	0.984	0.984	0.042	0.079	2894382.956	2894667.482
Indiana	CI	10433	2056.374	275		0.992	0.991	0.025	0.040	1928420.139	1928964.094
	WI		2296.707	291	240.333	0.991	0.991	0.026	0.080	1928628.472	1929056.383
	SI		4603.455	307	2306.748	0.981	0.981	0.037	0.088	1930903.220	1931215.088
Kansas	CI	8894	2110.578	275		0.991	0.990	0.027	0.043	1634280.583	1634812.568
	WI		2524.264	291	413.686	0.988	0.988	0.029	0.099	1634662.268	1635080.763
	SI		4233.895	307	1709.631	0.980	0.980	0.038	0.099	1636339.899	1636644.904
Kentucky	CI	6520	1179.054	275		0.994	0.994	0.022	0.029	1206322.986	1206831.683
	WI		1234.003	291	54.949	0.994	0.994	0.022	0.049	1206345.935	1206746.110
	SI		2077.320	307	843.317	0.989	0.989	0.030	0.058	1207157.252	1207448.905
Michigan	CI	6833	1299.092	170		0.992	0.991	0.031	0.052	989251.073	989660.844
	WI		1394.311	182	95.219	0.991	0.991	0.031	0.086	989322.292	989650.109
	SI		2839.156	194	1444.845	0.981	0.981	0.045	0.092	990743.137	990989.000
Minnesota	CI	19994	3713.725	170		0.991	0.990	0.032	0.043	2878450.276	2878924.468
	WI		3910.535	182	196.81	0.991	0.991	0.032	0.070	2878623.087	2879002.440
	SI		8089.829	194	4179.294	0.981	0.981	0.045	0.078	2882778.381	2883062.896
South Carolina	CI	20000	4609.603	90		0.988	0.986	0.050	0.076	2123841.525	2124197.182
	WI		4891.837	98	282.234	0.987	0.986	0.049	0.112	2124107.759	2124400.188
	SI		6308.980	106	1417.143	0.983	0.983	0.054	0.118	2125508.902	2125738.104
Washington	CI	12957	2033.200	275		0.995	0.994	0.022	0.026	2385763.775	2386323.979
	WI		2181.778	291	148.578	0.994	0.994	0.022	0.052	2385880.353	2386321.047
	SI		3608.474	307	1426.696	0.990	0.990	0.029	0.060	2387275.050	2387596.233
Wisconsin	CI	14926	2718.167	170		0.992	0.991	0.032	0.063	2148535.797	2148992.449
	WI		2901.072	182	182.905	0.991	0.991	0.032	0.093	2148694.703	2149060.024
	SI		5040.002	194	2138.93	0.984	0.984	0.041	0.099	2150809.632	2151083.623

* The levels of model constraints restricted to be equal across grades are: CI: Configural Invariance, WI: Weak Invariance, SI: Strong Invariance

TABLE 8

Goodness-of- Fit Indexes of Invariances of Linear MLGM Models of MAP Mathematics Tests

State	Model*	N	χ^2	df	$\Delta\chi^2$	CFI	TLI	RMSEA	SRMR	AIC	BIC
Colorado	CI	5417	3200.482	405		0.986	0.985	0.036	0.033	1202562.013	1203155.770
	WI		3327.537	425	127.055	0.985	0.985	0.036	0.057	1202649.064	1203110.875
	SI		5630.506	445	2302.969	0.974	0.974	0.046	0.065	1284353.034	1284814.845
Illinois	CI	20000	11337.522	275		0.982	0.981	0.045	0.029	3618796.729	3619389.491
	WI		12480.270	291	1142.748	0.980	0.980	0.046	0.082	3619907.480	3620373.786
	SI		20906.282	307	8426.012	0.967	0.968	0.058	0.088	3628301.492	3628641.342
Indiana	CI	10247	148965.438	560		0.580	0.553	0.161	3.559	2790203.181	2790962.829
	WI					**	**	**	**	**	**
	SI		135770.462	608		0.617	0.625	0.147	**	2776912.205	2777324.585
Kansas	CI	8966	4764.568	170		0.979	0.977	0.055	0.048	1264358.923	1264784.995
	WI		5255.879	182	491.311	0.977	0.976	0.056	0.121	1264826.234	1265167.091
	SI		9666.501	194	4410.622	0.957	0.958	0.074	0.126	1269212.855	1269468.498
Kentucky	CI	6602	3465.561	275		0.983	0.982	0.042	0.020	1211001.089	1211510.724
	WI		3652.142	291	186.581	0.982	0.982	0.042	0.059	1211155.670	1211556.582
	SI		6026.435	307	2374.293	0.970	0.970	0.053	0.070	1213497.964	1213790.154
Michigan	CI	6827	4184.200	275		0.981	0.979	0.046	0.024	1245720.842	1246232.990
	WI		4373.172	291	188.972	0.980	0.980	0.045	0.059	1245877.814	1246280.703
	SI		7279.799	307	2906.627	0.966	0.967	0.058	0.071	1248752.440	1249046.072
Minnesota	CI	20000	15870.194	170		0.970	0.966	0.068	0.093	2857014.335	2857488.544
	WI		17282.766	182	1412.572	0.967	0.966	0.069	0.139	2858402.908	2858782.275
	SI		27832.970	194	10550.204	0.947	0.948	0.084	0.142	2868929.112	2869213.637
South Carolina	CI	20000	13233.352	275		0.980	0.978	0.049	0.027	3649802.550	3650395.312
	WI		14894.473	291	1661.121	0.977	0.977	0.050	0.095	3651431.671	3651897.977
	SI		22617.917	307	7723.444	0.965	0.966	0.060	0.094	3659123.116	3659462.966
Washington	CI	14278	8230.010	170		0.978	0.976	0.058	0.046	2049743.767	2050197.755
	WI		9544.553	182	1314.543	0.975	0.974	0.060	0.125	2051034.309	2051397.500
	SI		14495.506	194	4950.953	0.962	0.962	0.072	0.135	2055961.262	2056233.656
Wisconsin	CI	15356	9389.812	275	9389.812	0.980	0.978	0.046	0.024	2780059.981	2780632.926
	WI		9653.384	291	263.572	0.979	0.978	0.046	0.047	2780291.554	2780742.270
	SI		13743.973	307	4090.589	0.970	0.971	0.053	0.055	2784350.143	2784678.631

* The levels of model constraints restricted to be equal across grades are: CI: Configural Invariance, WI: Weak Invariance, SI: Strong Invariance.
**: Not available.

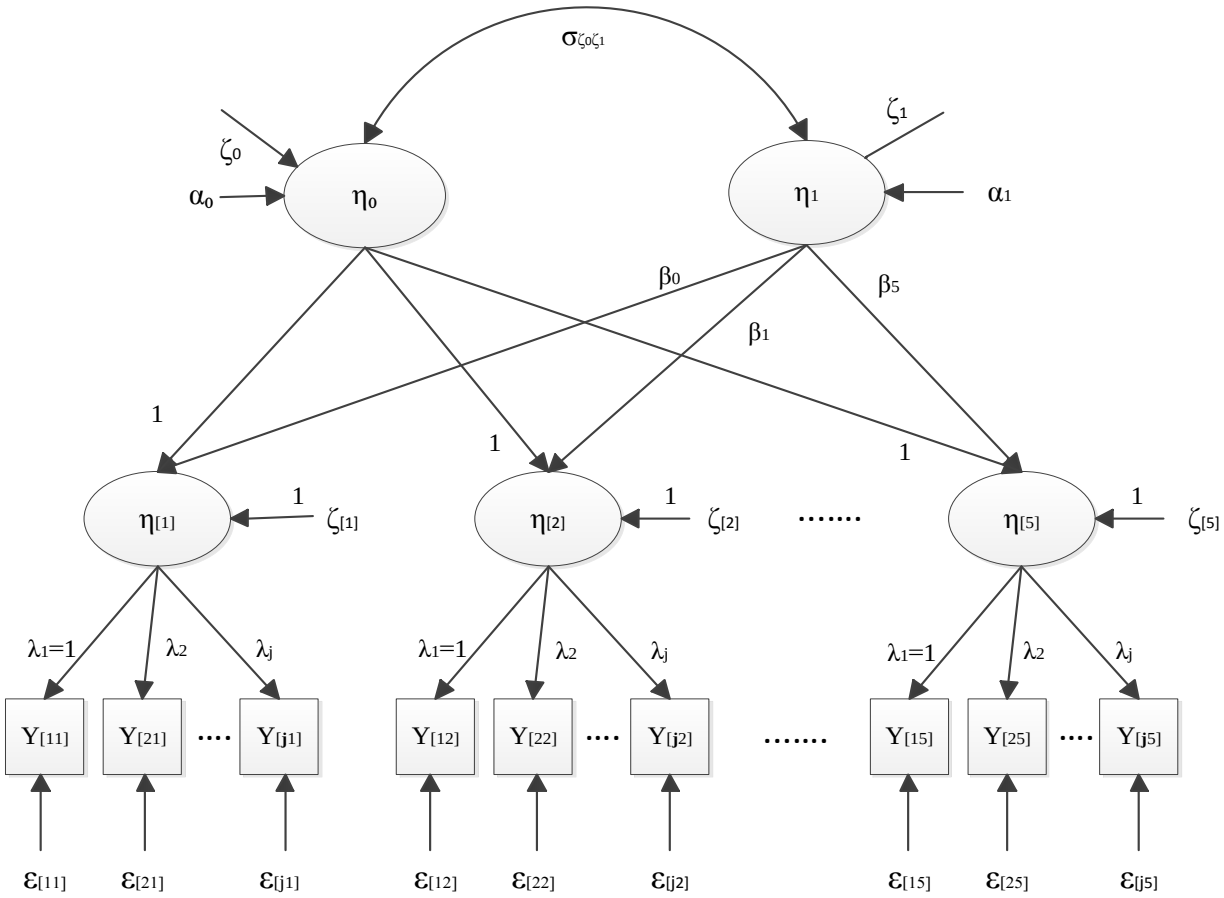


FIGURE 1. The MLGM at five time points with linear growth structure and invariance of factor loadings.

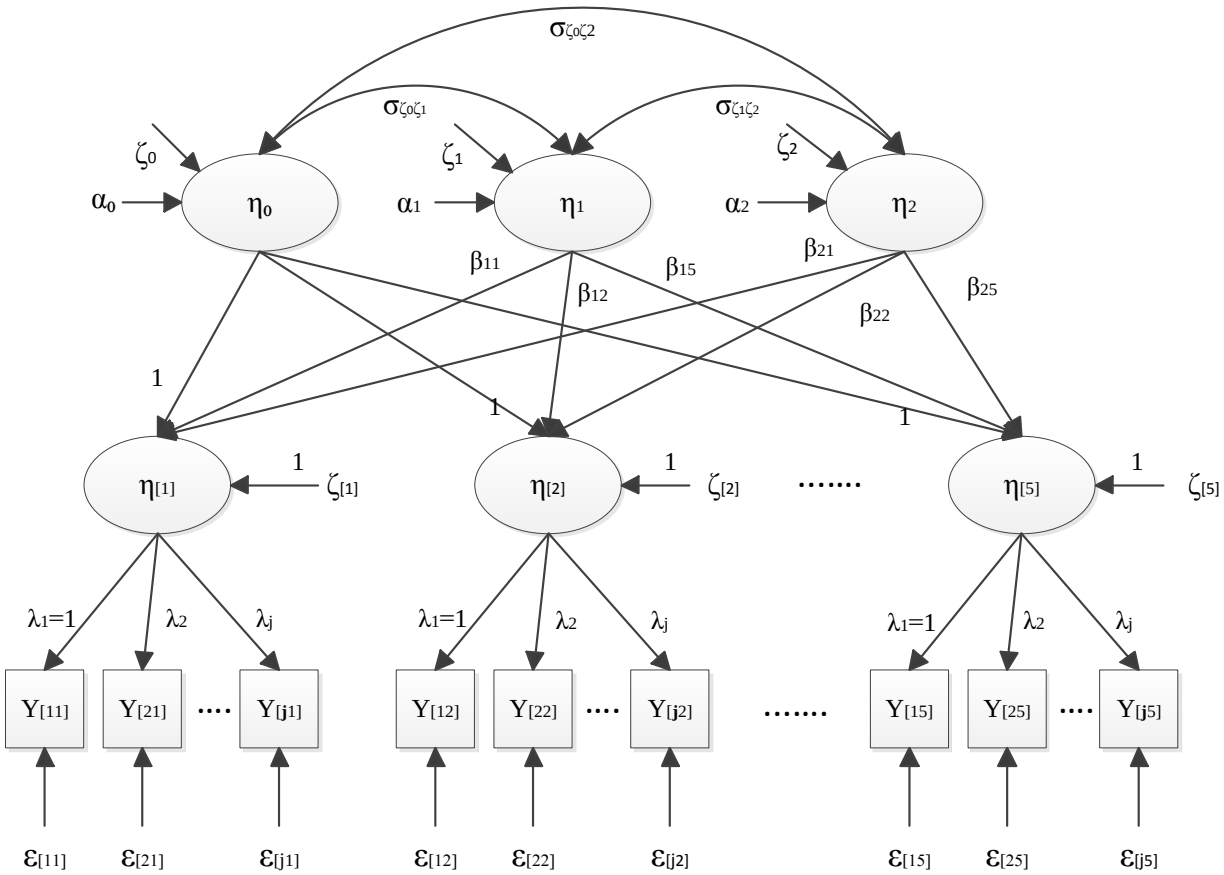


FIGURE 2. The MLGM at five time points with quadratic growth structure and invariance of factor loadings.

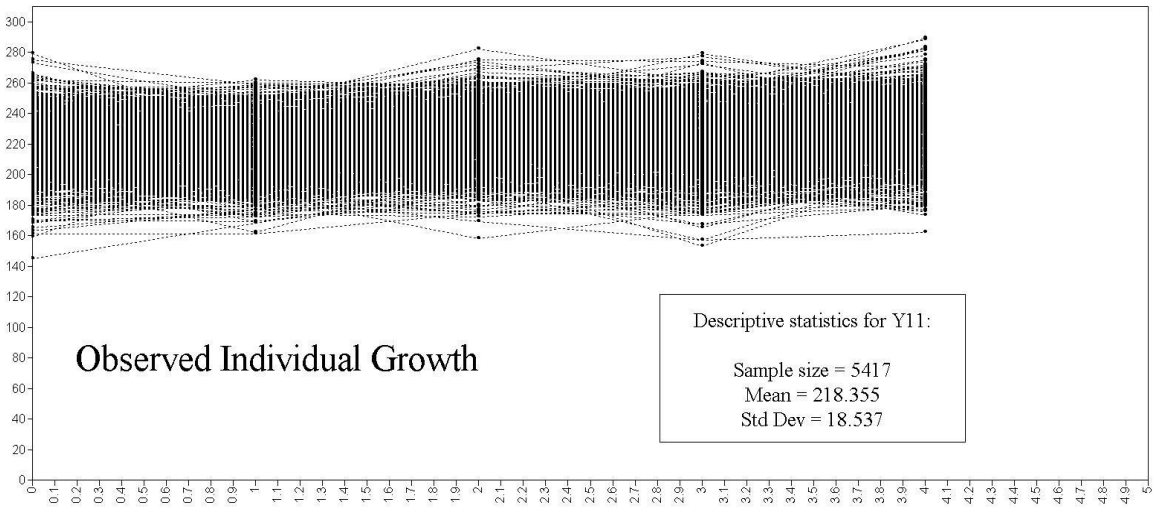


FIGURE 3. Observed individuals growth of $Y_{[11]}$ of MAP mathematics test across five-wave for Colorado

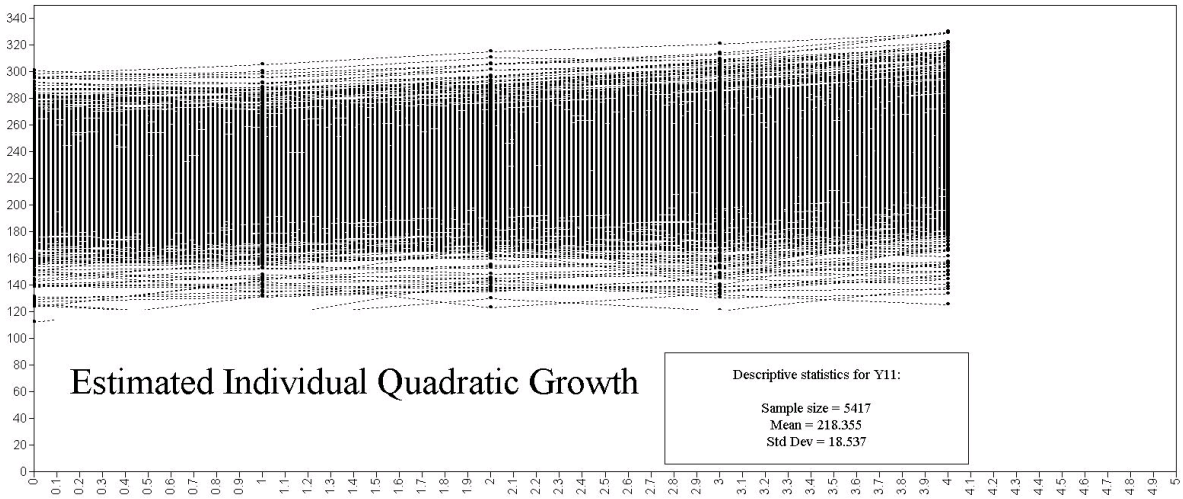


FIGURE 4. Estimated individuals quadratic growth of $Y_{[11]}$ of MAP mathematics test across five-wave for Colorado

REFERENCES

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika, 52*, 317–332.

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

Bandalos, D. L. (2002). The effects of item parceling on goodness-of-fit and parameter estimate bias in structural equation modeling. *Structural Equation Modeling, 9*, 78–102.

Bandalos, D. L., & Finney, S. J. (2001). Item parceling issues in structural equation modeling. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling: New developments and techniques.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Bergman, L. R., Eklund, G., & Magnusson, D. (1991). Studying individual development: Problems and methods. In D. Magnusson, L.R. Bergman, G. Rudinger, & B. TÄorestad (Eds.), *Matching problems and methods in longitudinal research* (pp. 1-28). Cambridge: Cambridge University Press.

Betebenner, D.A., & Linn, R. L. (2010). *Growth in student achievement: Issues of measurement, longitudinal data analysis and accountability*. Retrieved from http://www.k12center.org/publications.html.

Bollen, K. A. & Curran, P. J. (2006). *Latent curve models: A structural equation perspective*. Hoboken, NJ: Wiley.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9* 233-255.

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

Cronbach, L. J. (1988). Test validation. In R. L. Thorndike (Ed.), Five perspectives on validity argument. In H. Wainer & H. I. Braun (Eds), *Test validity* (pp-. 3-17). Hillsdale, N.J. : Brlbaum.

Cizek, G. J. (2005). Adapting testing technology to serve accountability aims: The case of vertically-moderated standard setting. *Applied Measurement in Education, 18*(1), 1-10.

Doran, H. C. & Cohen, J. (2005). The confounding effect of linking bias on gains estimated from value-added models. In R. W. Lissitz (ed.), *Value-added models in educations: Theory and applications (pp. 80-104*). Maple Grove, MN: JAM Press.

Drasgow, F. (1987). Study of the measurement bias of two standardized psychological tests. *Journal of Applied psychology, 72*:19-29.

Ferrer, E., Balluerka, N., & Widaman, K.F. (2008). Factorial Invariance and the specification of second-order latent growth models. *Methodology, 4*, 22-36.

Guilford, J. P. (1946). New standards for test evaluation. *Educational and Psychological Measurement, 6* 427-439.

Haertel, E., & Calfee, R. (1983). School achievement: Thinking about what to test. *Journal of Educational Measurement, 20*(2), 119-131.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-Stakes testing, *Educational Measurement Issues and Practice*, Vol: 23, Issue: 1.

Hall, R. J., Snell, A. F., & Foust, M. (1999). Item parceling strategies in SEM: Investigating the subtle effects of unmodeled secondary constructs. Organizational Research Methods, 2, 233–256.

Hamilton, L.S., Stecher, B. M., & Yuan, K. (2008). *Standards-based reform in the United States: History, research, and future directions*. RAND Corporation.

Hancock, G. R., Kuo, W., & Lawrence, F. R. (2001). An illustration of second-order latent growth models. *Structural equation Modeling, 8*(3): 470-489.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55.

Kingsbury, G. G., & Weiss, D. J. (1983). A comparison of IRT-based adaptive mastery testing and a sequential mastery testing procedure. In D. J. Weiss (Ed.*), New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 257-238). New York: Academic Press.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating: methods and practices* (2nd ed.). New York: Springer.

Li, Y. (2011). *Exploring the full-information bifactor model in vertical scaling with construct shift*. Unpublished doctoral dissertation, University of Maryland.

Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education, 6*, 83-102.

Linn, R. L. (2001). *The Design and Evaluation of Educational Assessment and Accountability Systems* (No. CSE Techinical Report 539). Los Angeles, CA: Center for the Study of Evaluation (CSE), National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Lissitz, R. L. (Ed.) (2006). *Longitudinal and value added models of student performance*. Maple Grove, MN: JAM Press.

Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002) To parcel or not to parcel: Exploring the question and weighing the merits. *Structural Equation Modeling, 9*, 151-173.

Loehin, J. C. (2004). *Latent variable models: An introduction to factor, path, and structural equation analysis*. Mahwah, NJ: Erlbaum.

Lohman, D. F. (1993). Teaching and testing to develop fluid abilities. *Educational Researcher, 22*(1), 12-23.

Lord, F. M. (1977). A broad-range tailored test of verbal ability. *Applied Psychological Measurement, 1*, 95-100.

Marsh, H. W., Hau, K. T., & Grayson, D. (2005). Goodness of fit in structural equation models. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Contemporary Psychometrics: A Festschrift for Roderick P. McDonald* (pp.225-340). Mahwah, NJ: Lawrence Erlbaum Associates.

Martineau, J. A., Paek, P., Keene, J., & Hirsch, T. (2007). Integrated, Comprehensive Alignment as a Foundation for Measuring Student Progress. *Educational Measurement: Issues & Practice. 26*(1), 28-35.

McArdle, J. J. (1988). Dynamic but Structural Equation Modeling of Repeated Measures Data. In J. R. Nesselroade & R. B. Cattell (Eds.), *The Handbook of Multivariate Experimental Psychology*. New York: Plenum Press.

McArdle, J.J., Grimm, K.J., Hamagami, F., Bowles, R.P., & Meredith, W. (2009). Modeling lifespan growth curves of cognition using longitudinal data with multiple samples and changing scales of measurement. *Psychological Methods, 14*, 126-149.

McDonald, R. P. (1985). *Factor analysis and related methods*. Hillsdale, N. J.: Erlbaum.

Meredith, W. M. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika, 58*, 525-543.

Messick, S. (1984). The psychology of educational measurement. *Journal of Educational measurement 21*, 215-237.

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

Messick, S. (1995). *Standards-based score interpretation: Establishing valid grounds for valid inferences*. In Proceedings of the joint conference on standard setting for large-scale assessments of the National Assessment Governing Board (NAGB) and the National Center for Education Statistics (NCES) (Vol. II, pp. 291-305). Washington, DC: National Assessment Governing Board and National Center for Education Statistics.

Mislevy, R. J. (1992). *Linking educational assessments: Concepts, issues, methods, and rospects.* Princeton, NJ: ETS Policy Information Center.

Muthén, B. O. (1991). Analysis of longitudinal data using latent variable models with varying parameters. In L. Collins & J. Horn (Eds.), *Best methods for the analysis of change. Recent advances, unanswered questions, future directions* (pp. 1-17). Washington D.C.: American Psychological Association.

Muthén, B. O. (1995). *Longitudinal studies of achievement growth using latent variable modeling*. Technical Report, University of California, Los Angeles.

Muthén, L. & Muthén, B. (2007). Mplus User's Guide (Fifth Edition ed.). Los Angeles, CA: Muthén & Muthén.

Northwest Evaluation Association. (2011, January). *Technical manual for Measure of Academic Progress & Measure of Academic Progress for Primary Grades*. Portland, Oregon.

Patz, J. R. (2007). *Vertical scaling in standards-based educational assessment and accountability systems*. Washington, DC: Council of Chief State School Officers.

Raftery, A. E. (1995). Bayesian model selection in social research (with discussion). In *Sociological Methodology 1995* (P. V. Marsden, ed.) 111–163. American Sociological Association, Washington, DC.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology, pp. 321-334 in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, IV. Berkeley, California: University of California Press. Available free from Project Euclid.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons, New York.

Sayer, A. G. & Cumsille, P.PE. (2001). Second-order latent growth models. In Collins, L. LM. & Sayer, A. G. (Eds). *New methods for the analysis of change*. Washington, D. C.: American Psychological Association.

Schwartz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*, 461–464.

Smith, R. L, & Yen, W. M. (2006). Models for evaluating grade-to-grade growth. In R. W. Lissitz (Ed.), *Longitudinal and value added modeling of student performance* (pp. 82-94). Maple Grove, MN: JAM Press.

Sternberg, R. J. (1998). Abilities are forms of developing expertise. *Educational researcher, 27*(3), 11-20.

Thissen, D. & Mislevy, R.J. (1990). Testing algorithms. In H. Wainer, N. Dorans, R. Flaugher, B. Green, R. Mislevy, L. Steinberg & D. Thissen, *Computerized adaptive testing: A primer*. Hillsdale, NJ: Lawrence Erlbaum Associates, 103-135.

Tisak, J., & Meredith, W. (1990). Descriptive and associative developmental models. In von Eye (Eds), *Statistical methods in longitudinal research*. Vol.2. Boston: Academic.

U.S. Department of Education (2009). Executive Summary, Race to the Top Assessment Program: Notice of public meetings and request for input. Retrieved from http://ed.gov/programs/racetothetop-assessment/executive-summary.pdf.

Wang, S., & Harris, G. (2011). *Psychometric evaluation of NWEA item calibration procedure*. Technical Report, Northwest Evaluation Association, Portland, Oregon.

Wang, S., & Jiao, H. (2009). Construct equivalence across grades in a vertical scale for a K-12 large-scale reading assessment. *Educational and Psychological Measurement, 69*(5), 760-777.

Wang, S., McCall, M., Jiao, H., & Harris, G. (2011). *Construct validity and measurement invariance of computerized adaptive testing: Application to Measures of Academic Progress (MAP) using confirmatory factor analysis*. Proposal for 2012 Annual Meeting of American Educational Research Association.

Way, W., Twing, J., Camara, W., Sweeney, K., Lazer, S., & Mazzeo, J. (2010). *Some considerations related to the use of adaptive testing for the Common Core Assessments*. Retrieved June 11, 2010, from www.ets.org/s/commonassessments/pdf/ AdaptiveTesting.pdf.

Widaman, K. F., & Reise, S. P. (1993). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In Bryant, K. J.; Windle, M.; West, S. G (Eds.). *The science of prevention: Methodological advances from alcohol and substance abuse research*. Washington, DC.: American psychological Association; p.281-324.

Williamson, G. L., Appelbaum, M., & Epanchin, A. (1991). Longitudinal Analyses of Academic Achievement. *Journal of Educational Measurement, 28*(1), 61-76.

Wise, L. (2004). Vertically-articulated content standards. Retrieved June 5, 2006, from http://www.nciea.org/publications/RILS_LW04.pdf.

Yen, W. M. (2007). Vertical scaling and No Child Left Behind. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 273–283). New York, NY: Springer.

Yen, W. M. (2009). *Growth models for the NCLB growth model pilot*. Princeton, NJ: ETS.

Turning the page: How smarter testing, vertical scales, and understanding of student engagement may improve our tests

G. Gage Kingsbury and Steven L. Wise
Northwest Evaluation Association

Building an adaptive testing system requires the practitioner to blend basic research (Weiss, 1980) with technology and logistics that fit the situations in which the tests will be delivered (Sands, Waters, and McBride, 1997). In this study, we will consider the research needs of an adaptive testing system that has been developed for use in a number of K-12 settings (MAP: Northwest Evaluation Association, 2009). While many of the operational questions associated with the initial development of this system have been described earlier (Kingsbury and Houser, 1993; 1998), this study focuses on issues that relate current research findings to ongoing adaptive test development and administration.

Primary and secondary educational systems bring unique issues to test development. Students commonly do not pay for the privilege of taking a test (as in certification, licensure, and college admissions testing). They may not be the focus of the test, and may not find out the result of their test (as in NAEP testing and some NCLB testing). Students may have varying degrees of motivation and preparation for a test. Students may not know how to read well, which may hamper testing of other subjects. Students range in age from about 5 to about 18 which provides all sort of interesting interactions between the student and the test.

The upshot of these unique characteristics of K-12 testing is that there are hundreds of issues to be considered in the development of an adaptive test to be used in schools. Many of these are logistic and practical issues (i.e., “How many items can a kindergartner

take in a test before losing focus?" or "How many tests can be given in a school day with a given number of computers?"). Others involve questions that have been addressed in the academic literature and can be applied with little change (i.e., "How do we select the next item to administer?" or "What scoring algorithm should we use?"). While these questions are crucially important for adaptive test development, we will not be dealing with them in this study.

Beyond these questions, there is a category of questions that need to be answered through research, and need to be answered within the context of K-12 education to be meaningful. We will discuss three of these questions in this chapter. The three questions that will be addressed are as follows:

1. Can a stable vertical measurement scale be established that allows the measurement of student achievement levels and achievement growth between grades and across years?
2. Can we identify students who are not giving full effort on a test, and how can we use the information if we can obtain it?
3. Can we adapt our tests using information aside from the momentary achievement estimate for the student and the psychometric characteristics of the items?

We have chosen to focus on these questions because they represent research questions in different states of done-ness. The question of vertical scales has been answered in some settings and the methodology exists to answer it in any setting. The question of student effort has been answered in some settings, but we know less about the processes for using the information. Finally, the question of using additional information to adapt our tests has been asked, but is far from an answer. These questions suggest the

range and the direction of future research that might be useful for additional development of K-12 adaptive tests.

The stability of vertical measurement scales

Current assessment regulations under No Child Left Behind require that tests be able to measure student proficiency at each grade, and allow the use of growth information in school accountability. Questions concerning the development of measurement scales that have meaning across grade levels arise directly from the nature of the educational process. As students move from one grade to the next the things they are taught change noticeably. Student capabilities change markedly, from learning numbers to learning number facts to using numbers to solve simple calculation problems to understanding algebra to creating geometric proofs to understanding the uses of calculus in our daily life. It does not seem possible that a single measurement scale can capture this wide variety of change in achievement.

On the other hand, it is the challenge of measurement to create scales that measure underlying traits in circumstances that vary. If we were talking about the measurement of temperature rather than student achievement, few folks would argue that temperature is not a coherent trait. However, when we examine the measurement of temperature, its characteristics are quite similar to those that we encounter in student achievement measurement.

In order to measure temperature near room temperature, we can use a variety of different instruments. In your home, you probably have alcohol thermometers, bimetal thermometers, electronic thermometers, and the little plastic device that pops up when your turkey is done cooking. Each of these devices measures a different kind of physical change in the characteristics of objects when they are heated or cooled (expansion,

differential expansion, differential resistance, and wax melting into liquid, for the thermometers mentioned). However, each device uses its information to make an inference about a single underlying trait, which may be measured on the Celsius scale (or the Fahrenheit scale, or a variety of other temperature scales).

As we move to temperatures close to absolute zero or temperatures in the core of a star, different instrumentation is required to measure temperature, but the measurement scales remain the same. Further, the physical characteristics of the object having its temperature measured may change noticeably as the temperature changes. For instance, water changes from ice, to liquid water, to steam as its temperature changes, but it is still clear that the attribute being measured is the same, even though the object being measured changes its characteristics markedly. Scales for temperature have been established, and endure unchanged across years.

So, considering student achievement, rather than temperature, the question becomes whether we can create a scale that measures a useful characteristic of student achievement that has meaning, regardless of the items that are used to measure it or the activities the student uses to express it. Research indicates that we can create such a scale, and that the scale can maintain its measurement characteristics over long periods of time.

Creating a vertical measurement scale

Education is all about time. A student grows from a novice reader to an expert over a period of a few years. A school district changes its curriculum and its students begin to learn more quickly. A state has a change in its tax revenue, and schools shorten the school year, causing students to learn less. In order to make these effects on learning knowable, we need a measurement scales that span both grades and years. Changing tests and

fluctuating statistics have made it difficult for educators to find constant scales in the past.

This ability to measure consistently across time is never more important than within the context of an adaptive test being used across grades in a course of instruction. Each test that is given depends on the measurement scale to provide equivalence with other tests, and this need expands quickly when tests are given across a period of several years using an item pool that may also change across time.

Two measurement scales developed with these needs in mind are the reading and mathematics scales developed by the Northwest Evaluation Association. These scales, known as the RIT scales, are associated with large item banks that are used to develop achievement tests for use in many schools. The one-parameter logistic (1PL) IRT model (Wright, 1977), along with a sophisticated item linking design, was used to create these measurement scales during the late 1970s.

Since then, thousands of items have been added to these item banks. Each new item has been connected to the original measurement scale through the use of IRT procedures and systematic calibration design (Ingebo, 1997).

These measurement scales are used to develop adaptive tests and to measure individual student growth. Since both of these activities depend to a great extent on the item parameter estimates, it is crucial that the invariance assumption hold in this application. While some variability in item parameter estimates is expected, too much variability could cause growth measurement to be quite problematic. Growth is a difficult quantity to measure under the best conditions, so a stable scale is a prerequisite to maintaining accuracy in growth measures.

The stability of a vertical measurement scale

For purposes of this discussion, we can define scale stability as the ability of the scale to ascribe the same measure to the same amount of the attribute of interest over a period of time. This definition allows the characteristics of the tested population to change over time, while requiring the scale to give the same score (within measurement error) to individuals with the same capabilities taking tests at different times. Ideally, the scale would remain a consistent indicator across long periods of time, so that trends in the attribute of interest can be observed over many years, and so that change in the attribute within a single individual may be observed.

Even though long-term scale stability is imperative to our ability to observe patterns of growth, few studies have examined the long-term stability of IRT item parameter estimates. Two studies that have investigated the issue were conducted by Bock, Muraki, and Pfeiffenberger (1988) and by Sykes and Fitzpatrick (1992).

Bock et al. (1988) investigated the stability of the item parameter estimates in the 3-parameter logistic IRT model from the College Board Physics Achievement Test over a period of ten years using an ANOVA design and looking for a two-way interaction between items and occasions. The authors found that there was a statistically significant drift in item difficulty across time. The authors interpreted the drift as being due to changes in physics instruction across the time period under investigation. The authors performed a similar analysis of the College Board English Achievement Test, and found no evidence of drift. Since the focus of this study was on the development of a statistical model to allow for drift, rather than on the drift itself, the authors did not discuss the impact of the observed drift on test scores.

Sykes and Fitzpatrick (1992) investigated the stability of 1-parameter logistic item parameter estimates for 285 items from a professional licensure test administered over a period of five years. This study found drift in item difficulty parameter estimates that was directional, with items being estimated to be more difficult across time. When the investigators examined the source of the drift, it did not seem to be associated with item position or item type. As in the previous study, the authors hypothesized that the change in difficulty estimates was associated with changes in curricular emphasis. Since the emphasis in this study was on the covariates of drift, the authors didn't discuss the impact of drift on candidate scores.

In practice, IRT item parameter *estimates* will not be invariant. Estimates will vary due to a number of factors that have been researched fairly extensively in the past. These factors include sampling fluctuation (Swaminathan & Gifford, 1983), departures from unidimensionality (Bejar, 1980), and other characteristics of the calibration design such as item context (Yen, 1980). To make matters more complex, the type of test used to create item parameter estimates and the algorithm used to compute the estimates will also influence the stability of the item parameter estimates (Ban, et al, 2001). All of these factors that may affect the accuracy of item parameter estimates suggest that we should be cautious in relying on the invariance property of IRT in practical settings without verification.

While the research in the literature is quite useful, it is practical to describe in detail a study done with the vertical scales that are used with the MAP (Measures of Academic Progress) tests. This study, reported by Kingsbury (2003), extends the earlier work in several ways. First, it investigates stability of item parameter estimates in a large item bank rather than a set of items used in a single test. Second, it uses a measurement

scale that has been designed to measure student growth across time, rather than tests designed to be taken only once. Third, it uses a longer elapsed time since initial calibration, ranging from 7 to 22 years. Fourth, it examines the amount of impact that item parameter drift might have on student scores. The full study included information about both reading and mathematics, but only mathematics will be discussed here.

The stability of measurement scales may be viewed in two ways. First, we may investigate whether individual items have changes in their difficulty estimates across time. If there is more change (drift) than expected due to sampling variability, we may identify this as a problem with the invariance assumption. Second, we may ask what impact any identified drift may have on the test scores from our assessments and what impact the drift may have on decisions that are made as a result of the assessments. The Kingsbury (2003) study investigated the issue from both perspectives.

Method.

3091 mathematics items were administered to students from grades 2 to 10 in 10 school districts from 7 different states as a part of their districtwide assessment programs in the 1999-2000 school year. Each student took approximately 50 items. All items were multiple-choice, with original item difficulty estimates that were obtained at least 7 years prior to the study. Approximately 320 test forms were used in the study. Over 100,000 student test events were used for the study.

The items were administered within the context of an achievement level test (Kingsbury & Houser, 1997). An achievement level test is a paper-and-pencil test that has approximately seven different forms (levels) designed to differ in difficulty. Students are administered a particular form chosen for them individually based on past tests scores or

using scores from a routing test. This design is equivalent to a two-stage adaptive test (Lord, 1971).

Any individual student took approximately 50 mathematics items. Since different achievement level tests were used in the different school districts involved in this study, any individual item was seen by only a small sample of the students involved. The combination of all test forms across all school districts and grades resulted in the sparse data matrix that was used for calibrating all of the items in the study.

The original IRT item difficulty estimates for all of these items were created between 1977 and 1993. The mean time between the original calibration and the new calibration was 16 years and 1 month. The original item difficulty estimates were obtained using a marginal maximum-likelihood calibration procedure (Houser, Hathaway, & Ingebo, 1983).

The new item difficulty estimates were created using the data collected in the 1999-2000 school year. Since few students took the same items and no student took a very large percentage of the items, the calibration procedure used was a procedure designed for use with adaptive tests and other sparse data structures (Houser, Kingsbury, & Harris, 1997). This procedure is mathematically equivalent to the marginal maximum-likelihood calibration procedure originally used to calibrate the items.

After elimination of items with very small samples, 2359 items were available for use. Calibration sample sizes for these items ranged from 300 students to over 10,000 students. A minimum student sample size of 300 was established to correspond to the minimum sample size that was allowed in the original calibration procedure.

Analysis.

While there are a variety of statistical tools available for identifying parameter estimate drift (see Donoghue & Isham, 1998), the use of the 1PL model simplifies matters substantially. In this study, simple differences between original and new difficulty estimates were used. Two analyses were conducted for each measurement scale in the study.

Scale drift analysis. The scale drift analysis included several aspects. First, correlations between the new and original item difficulty estimates were calculated and compared to correlations seen in other studies using the same measurement scales. Next, frequency distributions of the differences between the original item difficulty estimates and the new item difficulty estimates were calculated. These allowed the examination of the variability in parameter estimates. Bias and mean absolute differences were also calculated and compared to standard deviations of student performance to begin to identify the impact of parameter drift. Finally, item parameter estimate differences were examined as a function of the original calibration date to identify whether the elapsed time between the two calibrations contributed to observed drift. This last analysis used a subset of the available items (2204 items) because some items were originally calibrated across several testing seasons.

Impact analysis. A second method of analyzing the effect of change in calibrations over time is to ask whether that change has a noticeable impact on students' scores. In this analysis two representative test forms used in the study were chosen as example tests. The two forms were middle-difficulty forms used in the fifth grade in a suburban school district in Indiana. For each of these forms two raw-score-to-RIT scoring tables were created, one using the original parameter estimates and one using the new item parameter estimates (for the 1PL IRT model, a particular number-correct score is associated with a

single scale score, dependent only on the item parameter estimates). The two scoring tables were then compared to identify the maximum difference caused by using the new item parameter estimates. By comparing the scale scores obtained from the two sets of calibrations for a particular raw score, we can identify how much a particular student's test score would have changed as a result of any item parameter drift that occurred.

Results

The observed correlation between the original and new item difficulties estimates was .967. While this correlation is close to unity, it is useful to compare it to correlations obtained in other studies using the same measurement scales. Ingebo (1997) described a series of experiments from the 1970s in which multiple, concurrent samples were drawn to calibrate a set of items from these scales, to identify the consistency of the calibrations. In those studies, the correlations of mathematics difficulty estimates across samples ranged from .95 to .99. The results from the current study mimic those from these earlier studies in which the samples were drawn concurrently.

Although the correlations provide some evidence of stability, they don't provide information about the differences observed on an item-by-item level. Figure 1 shows the frequency distribution of the differences observed subtracting the new calibration of item difficulty from the original calibration for each item. It can be seen that the distribution is fairly symmetric around a difference of zero. It can further be seen that few items have difficulty differences of more than ten RIT points (approximately one θ unit).

[Insert Figure 1 here]

Figure 2 shows the relationship between the original item difficulty estimates and the new difficulty estimates for each item. The relationship appears visually linear, and corresponds well to the superimposed line of identity. The figure shows no evidence of drift associated with the difficulty of the items, and gives no indication of a non-linear trend in the item calibrations.

[Insert Figure 2 here]

The directional drift (bias or average difference) in item difficulty estimates observed was -.17 RIT points in mathematics. To put this in context, the standard deviation of students' scores in sixth grade in the most recent norming study done using these measurement scales (NWEA, 2005) was 14.8 RIT points in mathematics. The drift that has occurred in the scale over the 16.1 years of elapsed time in the studied interval has had an impact of approximately .01 standard deviations on the mean item difficulty estimate.

The average absolute difference in parameter estimates was 4.53 RIT points, with a median absolute difference of 4 RIT points. As expected, this difference was larger than the directional drift, but still less than one-third of a standard deviation. Given the small values for directional drift, we would expect these differences to balance out in a test of reasonable length. This assumption will be investigated more completely in the impact analysis below.

Impact Analysis

Figures 3 shows the RIT scores obtained from each number correct score for original and new difficulty estimates. The figure shows clearly that the scores obtained from the new item difficulty estimates and the old estimates are extremely similar. For the mathematics test, the maximum difference that could occur was 1.1 RIT points and the average magnitude of difference was less than .5 RIT points. Since the smallest observable difference between two RIT scores is 1 RIT point, these differences are small enough to be rarely observable. Given a typical distribution of RIT scores, the difference would be less than 1 RIT point in 99 of 100 cases. Since the standard error of a score on one or these tests would be approximately 4 RIT points, the impact of the change in calibrations would not be expected to change instructional decisions.

[Insert Figure 3 here]

From this examination of the stability of a vertical measurement scale, we can draw two conclusions, as follows:

1) There was no substantial drift in item difficulty estimates across the timeframe of this study, and no trend was seen in changes in difficulty estimates as a function of time since initial calibration.
2) The largest observed change in student scores moving from the original calibrations to the new calibrations was 1.1 RIT points, with over 99% of expected changes being less that 1 RIT point.

While the overall conclusion of the study is that the measurement scales examined are stable across time, some individual items fluctuated noticeably from their original calibrations. This suggests the need for ongoing calibration analysis. Even with a fairly stable scale, individual items may have difficulties that vary across time. A follow on study will investigate the characteristics of these highly variable items. This study should enable us to identify whether the large changes in difficulty for a small number of items are possibly due to specific features of certain items or whether they might be due to changes in instruction that have reduced (or increased) a student's opportunity to learn the content in the question. Examples of items that might experience such fluctuation include the following:

- An example of change specific to the item would be an item asking for a definition of the word "radical" which has had three most common definitions since the early 1980s.
- An example of change related to opportunity to learn might be seen in an item asking about the characteristics of a retro-virus. In the late 1970's only college-level biology students would have been introduced to the concept, but now it is standard content in most high school biology courses.

Building and maintaining a stable measurement scale is as much an exercise in engineering as it is an exercise in calibration. The measurement scales under consideration here were originally designed using a four-square design (Wright, 1977) with multiple cross links within and across student grades. It is expected that this original development has contributed to the ongoing stability of the measurement scales studied here. Therefore,

while this study has indicated that stable measurement scales can be created in practice, it does not suggest that the use of IRT calibration alone will assure scale stability.

In public education, there is ongoing debate about the quality of schools. One overlooked element that causes the debate to continue is the inconsistent nature of much achievement information. Different tests are used to measure student achievement in different grades in many locations. If there isn't a consistent measurement scale linking these tests, comparison of performance across grades is difficult. In many cases, score equating is used to allow comparison from one year to another. While this is a useful statistical technique, it isn't designed to create stable measurement scales.

A stable measurement scale allows the development of curriculum-referenced interpretation of test scores. For instance, with the mathematics RIT scale, a student who was able to complete two-digit addition question correctly would obtain approximately the same score in the year 2002 that they would have obtained in 1980. With this development, changes in test scores can be related directly to changes in student capabilities. In turn, this will allow the identification of positive and negative trends in education as they happen.

The procedures used in creating the measurement scales examined in this study have been successful in creating stability. This is a requirement for good measurement and even more important if we are planning to measure change in a school or a nation across time. The results of this study indicate that we can create measurement scales that are meaningful not only for short-term comparisons, but for long-term studies as well.

Identifying students who are not trying their hardest

Our conception of validity as the gathering of a body of evidentiary support for the use of test scores has solidified over the last three decades (Messick, 1980; Kane, 2001, Shepard, 1997). Most of this evidence comes in the form of group statistics including correlations with important external variables, decision accuracy in groups and subgroups, and other information at the group level. If a score-based inference is to be about individual students, however, evidence at the group level is a necessary but not sufficient condition to validate the use of a test score for a particular purpose for a specific student (Kingsbury & Hauser, 2007).

What other evidence is needed? Construct-irrelevant factors (Haladyna & Downing, 2004) may influence the validity of some scores in a group, but not others. For example, if Jonathan, who is taking a math test, suffers from test anxiety to the degree that his score is adversely affected, test anxiety can be considered a construct-irrelevant factor influencing Jonathan's score. In contrast, if no one else in Jonathan's class suffers from test anxiety, test anxiety is not a construct-irrelevant factor influencing John's classmates' scores. This implies that some scores are likely to be less valid than others. This idea of focusing validation down to the level of specific scores is called *individual score validity* (ISV). One definition of ISV is the degree to which a particular score is free from construct irrelevant factors (Wise, Kingsbury, & Hauser, 2009).

In K-12 testing, a major ISV threat is student test-taking effort. That is, obtaining a valid score for a given student requires both (a) that the test being given is appropriate for estimating that student's proficiency level, and (b) the student devotes sufficient effort to

that test. Without adequate effort, a student's test performance is apt to underestimate what she knows and can do.

The problems posed by low student effort affect our organization in several ways. First, because the scores from our MAP tests are used by teachers to help them make instructional decisions for each of their students, a non-effortful score is misleading. If Katie doesn't give good effort to her reading test, and her score is consequently lower than it should be, she may receive instruction in reading she doesn't need. Second, because MAP is an interim test used to measure student growth (as the difference between consecutive RIT scores) over multiple time periods, each non-effortful test score messes up two growth scores. When the non-effortful score is the first score in the growth calculation, the "growth" score will be inordinately high. When it is the second score, the "growth" score may be negative. In either instance, growth can be seriously mis-estimated for that student. Worse, the presence of such odd-looking growth scores potentially (and unfairly) threatens the credibility of the MAP assessment. Finally, the presence of data from non-effortful test events threatens the validity of our test norms and may distort the calibrations of our embedded field test items.

Measuring Student Effort

It is therefore important to be able to identify test scores that have been distorted by low student effort. How are these identifications to be made with an adaptive test? One method might to be to simply ask students at the end of their test the degree to which they tried. While this type of self-report approach has some utility, it is unclear how truthfully students will respond, particularly if they are concerned about punishment (or at least disapproval) from their teachers if they indicated they didn't try on their test. A second, more data-based approach would be to compute person-fit indices for each

student. In these methods, the aberrance of the item response pattern is evaluated by comparing the correctness of each item response to the probability of a correct response according to the IRT model being used. Unfortunately, although such indices have utility with fixed item tests, they are not well suited for use with adaptive tests. In these types of tests, the CAT algorithm strives to administer items whose probability of being passed by the student is around .50. In these instances, both correct and incorrect responses would be equally aberrant, which undermines the sensitivity of the person fit index to detect aberrant response behavior. An additional drawback of both self-report measures and person fit indices is that they both consider the test event as a whole, which makes it difficult to detect changes in student effort during the course of the test event (Wise & Kong, 2005).

A third method for measuring student effort is provided by item response time, which can be readily collected during a CAT. Wise and Kong (2005) showed that response time can be used to differentiate responses that occurred very quickly (termed rapid-guessing behaviors) from the remaining responses (termed solution behaviors). They developed an index of student test-taking effort, called *response time effort* (RTE), which equals the proportion a student's responses that were solution behaviors. Wise and Kong show validity evidence for RTE as a measure of test-taking effort.

An additional source of information regarding student effort in an adaptive test is provided by rate at which items are passed (i.e., accuracy). Students who are trying hard on a CAT should pass roughly half of their items. Students who are not giving good effort, in contrast, should pass items at a rate that more resembles chance responding. Hence, for example, a student who passed less than 30% of the administered items might be presumed to have behaved in a non-effortful fashion during the test.

Based on RTE and accuracy information, Wise, Kingsbury et al. (2009) developed five flagging criteria for identifying MAP test events that indicated low ISV. The percentages of test events that were flagged during fall and spring testing were 5.8% and 6.6%, respectively. After removal of these test events from the data, the correlation between the fall and spring RIT scores increased from .79 to .82, even though the standard deviations of the RIT scores decreased in both testing seasons. In addition, removal of the test events accounted for most of the extreme positive and negative fall-to-spring growth scores. These findings are consistent with a conclusion that a source of construct-irrelevant variance had been removed.

The use of heuristic flags to identify unusual response patterns and response times is very practical, but it also requires new development when applying the heuristics to a new test or testing population. In order to create a test that was designed to be repeatable in a wide variety of testing situations, Hauser and Kingsbury (2010) developed the Wariness Index (WI), which was designed to identify whether item response times and item response correctness are in keeping with expectation for each person/item interaction. An initial study concerning the use of the wariness index indicated that it could be used to identify unusual performance in an adaptive mathematics test administered to a large sample of fourth grade students.

Correlates of Rapid-guessing Behavior

Three student characteristics have been identified that are related to test-taking effort. The most consistent finding has been a gender effect; males have a tendency to exhibit lower mean levels of effort (Eklöf, 2007; Wise & DeMars, 2010; Wise, Kingsbury, Thomason, & Kong, 2004; Wise, Ma, Kingsbury, & Hauser, 2010; Wise, Pastor, & Kong, 2009). In addition, one study found that lower-ability college students (as measured by

SAT scores) were more likely than higher-ability students to exhibit non-effortful test-taking behaviors (Wise, Pastor, et al., 2009). This finding should be interpreted cautiously, however, as there have been several studies in the same measurement setting that did not find such a relationship (Kong, Wise, Harmes, & Yang, 2006; Wise et al., 2006; Wise & DeMars, 2006; Wise & Kong, 2005). Finally, Wise et al., (2010), investigated MAP test events from grades 3-9, finding that, as grade increased, the occurrences of rapid-guessing increased.

A number of item characteristics have been found to be related to effort. Items that require more reading by the student are more likely to receive rapid-guessing behavior (Wise, 2006; Wise, Pastor, et al., 2009). Wolf, Smith and Birnbaum (1995) found that effort on test items was influenced by their levels of mental taxation (i.e., how much mental effort was required by an examinee to reach a correct answer), with more taxing items receiving less effort. Wise, Pastor, et al. (2009) found that items presenting a greater number of response options received fewer rapid guesses, as did items containing a graphic. Wise et al. (2010) found differences in rapid-guessing behavior across content area; students were substantially more likely to exhibit rapid guesses in reading than in math.

Regarding the context in which an item is presented, the position at which an item occurs in a test has been related to the amount of effort it receives. Items occurring later in a test are more likely to receive rapid guessing (Wise, 2006; Wise, Pastor, et al., 2009). Wise, Owens, Yang, Weiss, Kissel, Kong, and Horst (2005) found significant differences among test session proctors in the average amount of effort exhibited by examinees in their sessions. In addition, Wise et al., (2010) found that rapid guessing was related to the

time of day that testing occurred; as the day progressed through the morning into the afternoon, the more likely that rapid-guessing behavior was to occur.

Thus, the research to date has identified a number of correlates of rapid-guessing behavior. Characteristics of the students, the items, and the context in which the test is administered can each affect the likelihood that rapid guessing occurs. In the next section, we will examine how we might apply our knowledge of these correlates to improve student effort.

What can we do about low effort?

Having identified a student who did not try his or her hardest on a test, what can we do about it? There are several choices available. The first is simply to report it to the individual(s) who will make inferences about the student's score. Information regarding untrustworthy scores should be useful to anyone who interprets the scores, so developing reports that identify the scores reflecting low student effort will be useful. Second, because non-effortful behavior is usually identifiable in only some of a student's item responses, it may be possible to "salvage" a more valid score from those item responses that reflected solution behavior. Wise & DeMars (2006) developed an effort-moderated IRT model that effectively excludes rapid guesses from proficiency estimation. They found evidence of improvements in both model fit and convergent validity when this model was used, as compared to a traditional IRT model.

Although enhanced score reports and improved scoring models allow us to deal with non-effortful test taking after it occurs, there are more proactive options available. First, we might strategically test in ways that reduce how much rapid-guessing behavior occurs in the first place. One might, for example, test earlier in the day, using shorter tests with items that either do not involve a lot of reading or mental taxation, or use items that

include a table or graphic. A second option would be to try to identify rapid-guessing behavior as it occurs, and try to preempt it. Wise, Bhola and Yang (2006) developed an effort-monitoring computer-based test (CBT). In this type of test, the computer monitored item response times for rapid-guessing behavior, and displayed warning messages to students who had begun to exhibit non-effortful behavior. Two experimental studies (Kong, Wise, Harmes, & Yang, 2006; Wise, Bhola, et al., 2006) found that, for students deserving warnings, both effort and test performance were improved when an effort-monitoring CBT was used. The effort-monitoring CBT is a rudimentary example of a smarter adaptive test, which will be discussed in the next section.

Expanding the nature of test adaptation: Smart tests

The third aspect of adaptive testing that we will discuss here is the expansion of adaptive capabilities in testing to use additional information that can be made available during a test event. This development is in its infancy, but holds promise for improving testing practice in education and in other areas.

Most fixed-form tests used nothing more than a student's grade level to assign a test form that was designed in advance to match a content blueprint. As we have moved to adaptive testing, we began to use information about items (commonly IRT parameter estimates) to select items that were tailored to the performance of each student. Most of the techniques of item selection currently used in operational adaptive tests were developed in the 1970s and 1980s. It is time to begin to use the abundant additional information that we have in a testing situation to improve our tests.

In order to create smarter tests, we need to consider the elements in a testing situation that provide us information. Figure 4 shows some of the active elements present in most testing situations. The student brings his knowledge and achievement into a test environment that has a proctor, a test engine, and a host of other elements that might provide information about the student's capabilities and might also influence the outcomes from the test. While it is not a comprehensive list, here are some of the important influencers that the testing elements possess:

- Student – achievement, gender, involvement, interests, traits, states, fatigue, breakfast, educational history
- Environment – time of day time of year, lighting, noise, fire alarms
- Test Engine – user interface, delivery platform, test entry, item selection, scoring test termination, reporting, feedback, data capture, informational messages
- Proctor – training, experience, attitude, other jobs, interest, breakfast
- Hardware – backbone, bandwidth, local servers, video capabilities, audio quality
- Items – item pool characteristics, item difficulty, reading load, related items, multipart items, response format
- Test Purpose – stakes for student, stakes for teacher, stakes for school, stakes for external bodies, test length, use of test data in instruction

Our tests can become smarter through an understanding of most of these dimensions (including breakfast) but as examples, we will focus on student interests and on item selection.

Using student interests to make tests smarter

We know that students do better on tasks that involve topics in which they have a personal interest. We also know that students vary dramatically in their interests. We know that items reflect a wide variety of content that taps into different aspects of student interest. We also know that items can be created on the fly with known characteristics, including content designed to tap particular interests (Fredrickson, Mislevy, and Bejar, 1993). Given those pieces of information, it appears that we should be able to make our tests smarter by using information about a student's interests.

While there are a variety of ways to use information about student interests, some of the approaches might include the following:

- Select items with known non-trait-related content weighted by a student's interests
- Customize item non-trait-related content to adjust to the weighting of a student's interests
- Select items with modalities that match each student's learning styles
- Give students control over non-trait-related content in test questions

So, if we consider only the first approach (selecting items with particular content), an example might go as follows:

> Marie takes a short survey in class that indicates she is interested in horses and art. This information is captured and fed into the testing system. When Marie takes a test later in the semester, the system uses Marie's interests in choosing test questions. If, as Marie takes the test an item including content about horses or art is included in high-information items that might be chosen, the interest weighting

would increase the probability that Marie would see this question. In this manner, interest weighting would be used with algorithms that are currently used for exposure control, but in this case increasing the likelihood of seeing particular items for particular individuals. As a result, Marie would tend to see more items including horses and art than the average student, which may keep her more engaged in the testing experience.

While interest weighting has the potential to make our tests smarter through improving the match of content to interests, research is needed to determine whether this approach results in more meaningful test scores. Interest weighting could be combined with measures of student involvement, so that if a student was not engaged in the test, more items with content of interest would be administered. As we begin to make changes like this in our testing systems, it is important to continue to monitor scores for their predictive validity and generalizability.

Using new item selection techniques to make tests smarter

Current approaches to item selection use a limited set of item characteristics (IRT item parameter estimates, item content standard, etc.). These characteristics are used in the context of an existing test blueprint to create the final test of the required length with the required content distribution. For certification and licensure tests, this is sometimes modified to allow variable test lengths that result in decisions with a given confidence level. The most advanced approach currently used is a combination of Stocking's procedure for using multiple constraints (Stocking and Swanson, 1993) within the context of a shadow test (van der Linden and Veldkamp, 2004).

While these approaches provide tests of very high quality, we have more information about the items than we are currently using. Some of the information that we have about individual items includes reading load, non-trait related content, average response time, word count, response mode, depth of knowledge, and a host of other characteristics. These characteristics may be used to enhance our item selection. Some of these enhancements may include the following:

- controlling expected response time so that a test fits into a given time period, while still allow the test to be untimed
- adjusting the reading load of the test (in mathematics and science) to account for the reading achievement of the student
- individualizing the difficulty of items that a student sees to maintain maximal involvement
- individualizing weighting of items with different response modes to match the style of the student
- Controlling item selection to keep a student engaged

Once again, it is useful to give an example of the use of one of these approaches in selecting items for a student. Here we will consider controlling expected response times.

> Zack is a fourth grade student who is an able reader and pretty good in mathematics. As he is taking his smart adaptive test in math, he encounters a series of items that are challenging to him, but he perseveres and finishes each question. As he moves toward the end of the test, the item selection algorithm notices that

Zack seems to be losing his engagement in the test (using the tools described in the section above).

To reengage Zack, the smart test takes three different actions. First, it gives Zack an informative message suggesting that he continue to try his hardest. Second, it begins to select slightly easier questions for Zack, to allow him to reengage without much loss of accuracy. Third, it begins to overweight items which may be of interest to Zack even more than before. At the same time, the system alerts the proctor to watch Zack to make sure he isn't getting sick or in need of a bathroom break.

If Zack reengages in the test, we may be able to provide a score for Zack that is a more valid indicator of his achievement. This should have positive repercussions for Zack, his teachers, and our educational system as a whole.

Creating smarter test may allow us to have more control over test score validity but it is very much a work in progress. Whether it works in practice remains to be seen. Substantial research is needed in this area to identify whether and to what extent approaches to make tests smarter also improve the validity and generalizability of test scores without noticeably reducing score precision.

Ramifications

Stable vertical scales

In education, the impact of stable scales should be substantial. In other areas the development of stable scales has allowed us to measure length and weight and temperature. Any of these concepts can now be measured with consistent meaning from

one country to another and from one year to the next. There is no reason that educational achievement cannot be measured in the same manner, with consistent meaning and known precision.

The impact of vertical scales (when meshed with high-quality measurement instruments) may have educational impact that is more personal, but as substantial. Vertical scales allow the measurement of individual change as a student progresses from one year to the next and one classroom to another. Having this type of growth information at the student level will eventually allow us to examine how our educational system is serving each student.

In certification and licensure, the development of stable vertical scales opens up new possibilities for recertification and advanced licensing in the professions. We can imagine a time in which we can measure growth in professional capacity as well as initial competency, to provide a more complete picture of the capabilities of each candidate and professional.

Identification of student engagement

If a student is not attending to a test as well as they should, this might have a variety of interpretations, and it might mean

- that the student is not engaged in school
- that the student has difficulty with responsibility
- that the student is anxious in the testing setting
- that breakfast was bad

Some of these meanings are immediately addressable, but others require more long-term intervention. The initial indicator in the testing setting would suggest approaches to follow up to help the student thrive in education.

The evidence that indicates that individuals vary in their engagement with achievement tests is substantial. Further, the studies above show that we can identify many students who are not engaged in these tests. While the approaches that we use to identify the unengaged may differ, it is clear that identifying them is imperative. In the past, we have assumed that students are all equally engaged in using the trait of interest to answer the questions on each test that they take. If we are actually seeing results that differ do to differential engagement, we need to rethink many of our current conceptions of educational achievement.

As an example, we see differential test performance among groups, and we interpret it as an achievement gap. However, if the gap is in part caused by differential engagement, then our interpretation of that gap and corrective actions that we might take concerning the gap may change substantially. Only when we begin to measure student engagement can we expect to understand differences in average scores among groups. Beyond education, the techniques that are developed for identifying engagement in testing have immediate application to psychological testing. Much of the value of well-designed psychological instruments comes from their capability to identify individuals who are lying, responding in a socially desirable manner, or not taking the assessment seriously. The approaches designed to identify test engagement may allow us to obtain unobtrusive information that will predict these other psychological machinations more easily, and at less cost of individual time.

Creating smart tests

As we begin to create smarter tests, they should allow us to forge more direct relationships between assessment and instruction. The information from smart tests should allow us to provide teachers with information about individual response styles, speed of performance, appropriate level of challenge, and response to control. The smart test should allow a feedback loop to develop so that new information about the student from the teacher improves the next test the student takes, and new information from the test allows the teacher to understand how the student learns best.

Beyond education, the idea of smart tests should help change our psychological instruments, and may have substantial value in certification and licensure. In psychological testing, we can foresee a vocational instrument that adjusts as the assessment proceeds, so that the outcome can be much more tailored to the needs of the individual, combining abilities and strengths with interests and communication style. In certification testing, many aspects of job performance that aren't immediately apparent in a single test score can be captured in a smart test. The potential for substantial improvement in certification decisions helps not only the testing profession but the consumers of a wide variety of professional services as well.

Conclusion

We have considered stable scales, student engagement, and smarter tests. Research leading to testing decisions in these areas is at substantially different points of development. However, work in each of these areas is attempting to bring us to an era in

which our tests provide more information, more valid information, and more generalizable information.

In our application, decisions based on stable vertical scales have allowed us to determine how accurate our adaptive tests must be to allow us to measure individual growth during a year of instruction. They have also allowed us to connect test scores directly to the skill that a student knows now, what they are challenged by learning next, and what they aren't quite ready to learn yet. This information provides teachers with the ability to shape classroom instruction to match the needs of each of their students.

We are starting to implement processes to identify students who aren't engaged in the tests they are taking. We expect the ability to identify these students to have substantial impact in instruction, but we also expect it to change the conversation around effective schooling. Currently, low test scores are commonly viewed as indicating poor teaching, but engagement research may help us differentiate students who don't know the content from those who know the content but aren't engaged in the testing activity.

The next generation of computer-based tests promises to enhance our ability to obtain more valid scores. By expanding our conception of what an adaptive test can do, we will begin to develop smarter tests that can effectively manage construct-irrelevant factors that limit the validity of individual student scores.

References

Ban, J.-C., Hanson, B. A., Wang, T., Yi, Q., & Harris, D. J. (2001). A Comparative Study of On-Line Pretest Item: Calibration/Scaling Methods in Computerized Adaptive Testing. *Journal of Educational Measurement*, *38*, 191-212.

Bejar, I. I. (1980). A procedure for investigating the unidimensionality of achievement tests based on item parameter estimates. *Journal of Educational Measurement*, *17*, 283-296.

Bock, R. D., Muraki, E., & Pfeiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement*, *25*, 275-285.

Donoghue, J. R. & Isham, S. P. (1998). A comparison of procedures to detect item parameter drift. *Applied Psychological Measurement*, *22*, 33-51.

Eklöf, H. (2007). Test-taking motivation and mathematics performance in TIMSS 2003. *International Journal of Testing, 7,* 311-326.

Frederiksen, N., Mislevy, R. J., & Bejar, I. I. (1993) (Eds.). *Test theory for a new generation of tests*. Philadelphia, PA: Psychology Press.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23(1),* 17-27.

Hauser, C. & Kingsbury, G. G. (2010, June). In pursuit of individual score validity: The Wariness Index. Paper presented at the Rasch Conference: Probabilistic models for

measurement in education, psychology, social science, and health. Copenhagen, DK.

Houser, R. L., Kingsbury, G. G., & Harris, G. (1997). MATCAL: A program for calibrating items using data from sparse matrices. Portland, OR: NWEA.

Houser, R. L., Hathaway, W. E., & Ingebo, G. S. (1983). An alternate procedure to obtain ability estimates in latent trait models. Paper presented to the annual meeting of the American Educational Research Association, Montreal, Canada.

Ingebo, G. S. (1997). *Probability in the measure of achievement*. Chicago, IL: MESA Press.

Kane, M. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*(4), 319-342.

Kingsbury, G. G. (April, 2003). A Long-Term Study of the Stability of Item Parameter Estimates. Paper presented to the annual meeting of the American Educational Research Association, Chicago, IL

Kingsbury, G. G. & Hauser, C. (2007, April). Processes for Enhancing the Validity of Adaptive Test Scores. Paper presented to the annual meeting of the American Educational Research Association, Chicago, IL.

Kingsbury, G. G. & Houser, R. L. (1993). Assessing the utility of item response models: Computerized adaptive testing. *Educational Measurement: Issues and Practice*, 12-1, 21-27.

Kingsbury, G. G. & Houser, R. L. (1997). Using data from a level testing system to change a school district. In J. O'Reilly (Ed.), *The Rasch tiger ten years later: Using IRT techniques to measure achievement in schools* (pp. 10 - 24). Chicago, IL: National Association of Test Directors.

Kingsbury, G. G. & Houser, R. L. (1998). Developing computerized adaptive tests for school children. In Drasgow, F. and Olson-Buchanan, J. B. (Eds.) *Innovations in computerized assessment*. Mahwah, NJ: Lawrence Erlbaum Associates.

Kong, X. J., Wise, S. L., Harmes, J. C., & Yang, S. (2006, April). *Motivational effects of praise in response time-based feedback: A follow-up study of the effort-monitoring CBT*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Lord, F. M. (1971). A theoretical study of two-stage testing. *Psychometrika, 36,* 227-242.

Messick, S. (1980). Test validity and the ethics of assessment. American Psychologist, 35-11, 1012-1027.

Northwest Evaluation Association (2009). *Technical manual for Measures of Academic Progress and Measures of Academic Progress for Primary Grades.* Portland, OR: Author.

Sands, W., Waters, B., & McBride, J. (1997). *Computer adaptive testing: From inquiry to operation*. Washington, DC: American Psychological Association.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. Educational Measurement: Issues and Practice, 16-2, 5-8.

Stocking, M. L. & Swanson, L. (1993). A Method for Severely Constrained Item Selection in Adaptive Testing. *Applied Psychological Measurement, 17-3,* 277-292.

Swaminathan, H. & Gifford J. A. (1983). Estimation of parameters in the three parameter latent trait model. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing.* New York: Academic Press.

Sykes, R. C. & Fitzpatrick, A. R. (1992). The stability of IRT *b* values. *Journal of Educational Measurement*, *29*, 201-211.

van der Linden, W. J. & Veldkamp, B. P. (2004). Constraining Item Exposure in Computerized Adaptive Testing With Shadow Tests. *Journal of Educational and Behavioral Statistics, 29-3,* 273-291.

Weiss, D. J. (1980) (Ed.). New horizons in testing: Latent trait test theory and computerized adaptive testing. New York: Academic Press.

Wise, S. L. (2006). An investigation of the differential effort received by items on a low-stakes, computer-based test. *Applied Measurement in Education, 19,* 95-114.

Wise, S. L., Bhola, D., & Yang, S. (2006). Taking the time to improve the validity of low-stakes tests: The effort-monitoring CBT. *Educational Measurement: Issues and Practice 25(2),* 21-30.

Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement, 43,* 19-38.

Wise, S. L., & DeMars, C. E. (2010). Examinee non-effort and the validity of program assessment results. *Educational Assessment*, 15, 27-41.

Wise, S. L., Kingsbury, G. G., & Hauser, C. (2009, April). *How do I know that <u>this</u> score is valid? The case for assessing individual score validity*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego..

Wise, S. L., Kingsbury, G. G., Thomason, J., & Kong, X. (2004, April). *An investigation of motivation filtering in a statewide achievement testing program*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18,* 163-183.

Wise, S. L., Ma, L., Kingsbury, G. G., & Hauser, C. (2010, May). *An investigation of the relationship between time of testing and test-taking effort*. Paper presented at the annual meeting of the National Council on Measurement in Education, Denver, CO.

Wise, S. L., Owens, K. M., Yang, S., Weiss, B., Kissel, H. L., Kong, X., & Horst, S. J. (2005, April). *An investigation of the effects of self-adapted testing on examinee effort and performance in a low-stakes achievement test*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal.

Wise, S. L., Pastor, D. A., & Kong, X. J. (2009). Understanding correlates of rapid-guessing behavior in low stakes testing: Implications for test development and measurement practice. *Applied Measurement in Education, 22,* 185-205.

Wolf, L. F., Smith, J. K., & Birnbaum, M. E. (1995). Consequence of performance, test motivation, and mentally taxing items. *Applied Measurement in Education, 8*, 341-351.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14*, 97-116.

Yen, W. M. (1980). The extent, causes, and importance of context effects on item parameters for two latent trait models. *Journal of Educational Measurement*, *17*, 297-311.

Tables and Figures

Figure 1. Frequency of mathematics items as a function of the difference between the original item difficulty estimate and the new item difficulty estimate on the RIT scale (rounded to the nearest integer).

Figure 2. Relationship of original and new item difficulty estimates on the RIT scale for 2359 items in mathematics with superimposed identity line (r = .967).

Figure 3. RIT scores as a function of obtained number correct score calculated using original and new mathematics difficulty estimates.

Figure 4. Some of the primary elements that have an influence on the outcomes of a test event.

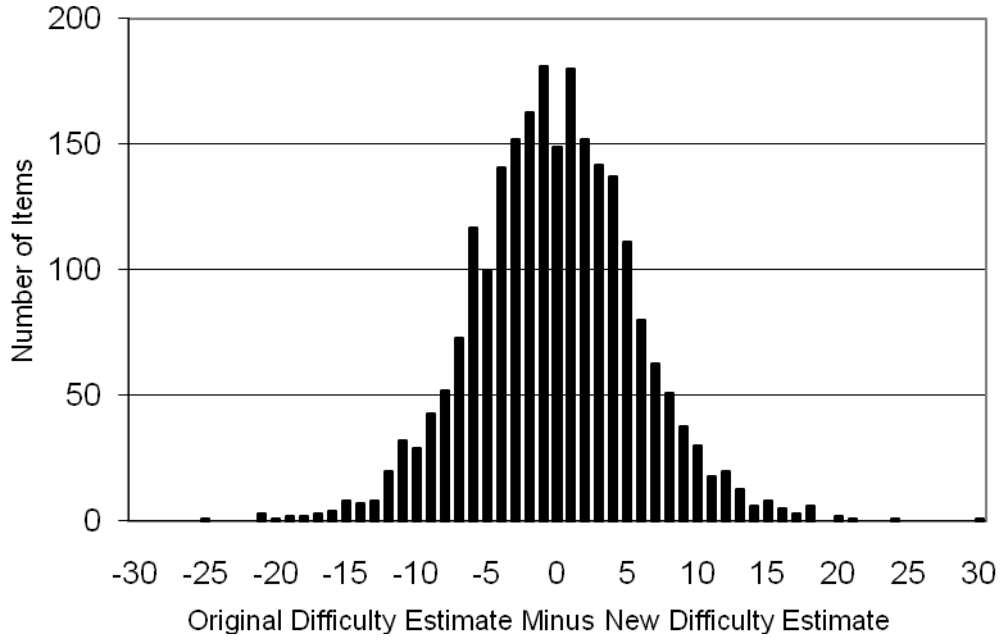


Figure 1. Frequency of mathematics items as a function of the difference between the original item difficulty estimate and the new item difficulty estimate on the RIT scale (rounded to the nearest integer).

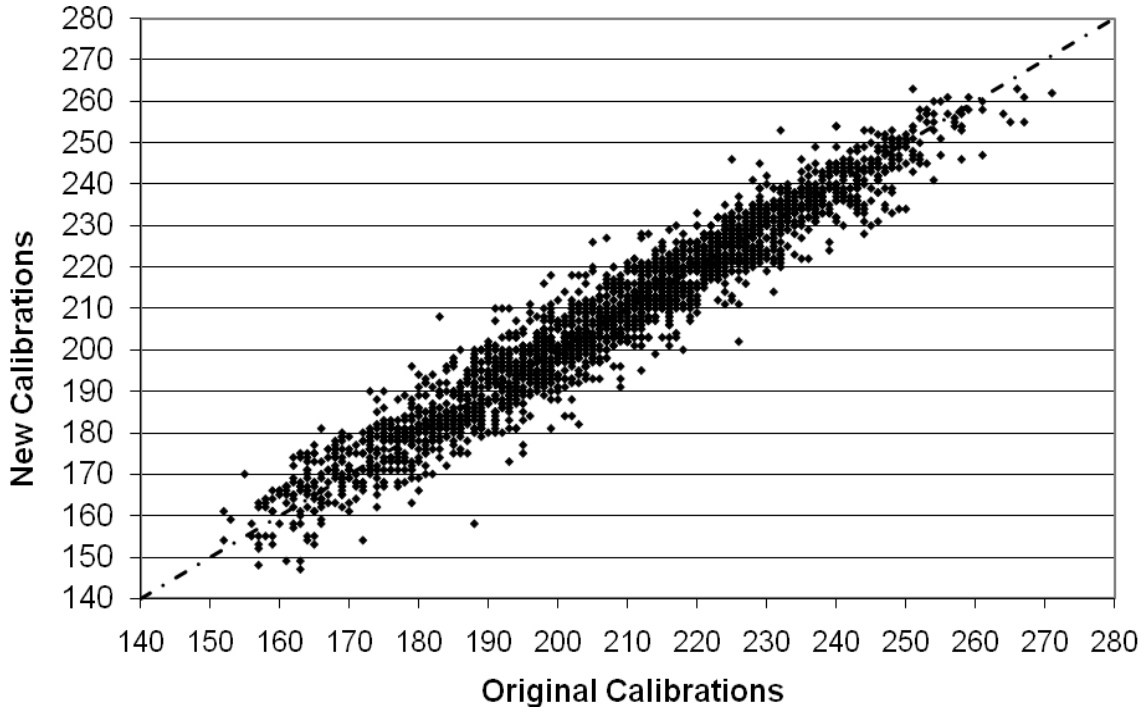


Figure 2. Relationship of original and new item difficulty estimates on the RIT scale for 2359 items in mathematics with superimposed identity line (r = .967).

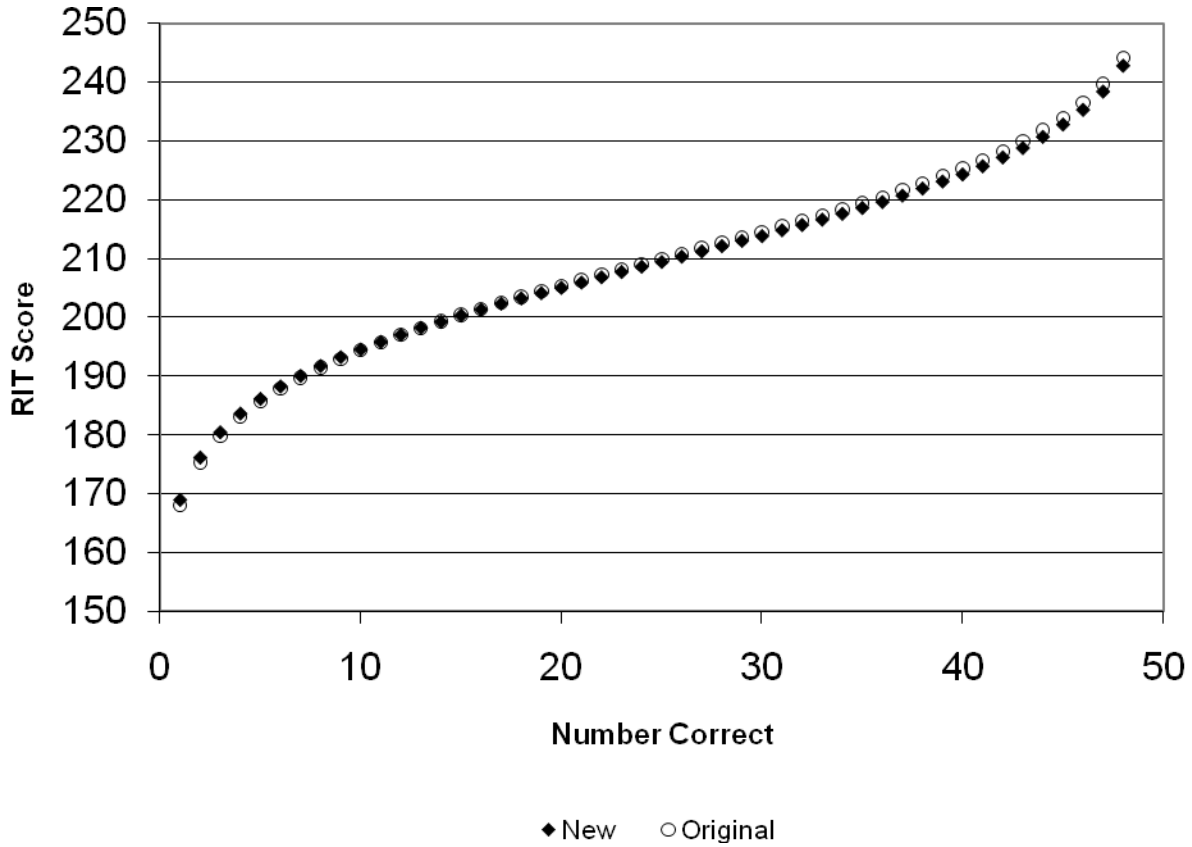


Figure 3. RIT scores as a function of obtained number correct score calculated using original and new mathematics difficulty estimates.

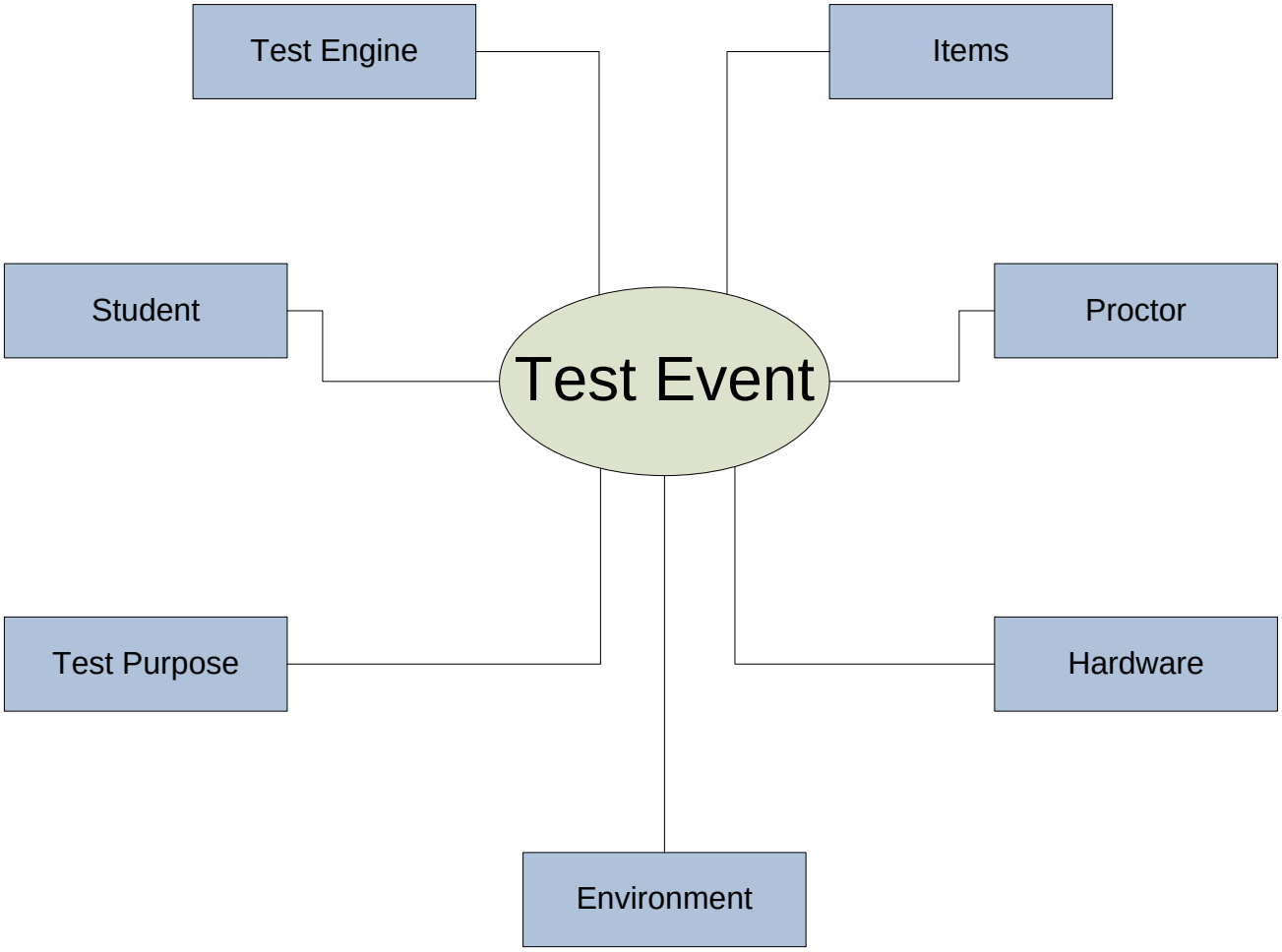


Figure 4. Aspects of the educational ecosystem that influence the interaction between a student and a test

The Utility of Adaptive Testing in Addressing the Problem of Unmotivated Examinees

Steven L. Wise

Northwest Evaluation Association

Author Note

Correspondence concerning this article should be addressed to Steven L. Wise, Northwest Evaluation Association, 121 NW Everett St., Portland, OR 97209. E-mail: steve.wise@nwea.org

Abstract

In low-stakes testing programs, test givers have the dual responsibilities of developing and administering a high-quality test that can yield valid scores and motivating examinees to give good effort to that test. However, unmotivated examinees represent a major threat to validity of scores from these types of tests. This integrative review examines the motivational benefits of computerized adaptive tests (CATs), and shows that they can have important advantages over conventional tests in both identifying instances when examinees are exhibiting low effort and effectively addressing the validity threat posed by unmotivated examinees.

The Utility of Adaptive Testing in Addressing the Problem of Unmotivated Examinees

When we administer cognitive tests, we seek valid information about what examinees know and can do. To that end, we typically encourage examinees to demonstrate the highest level of performance of which they are capable (Cronbach, 1960). It has long been recognized, however, that "unless people are motivated to do their best on a test, their scores will not reflect their maximum performance capabilities" (Betz & Weiss, 1976a, p. 1). The impact of test-taking motivation on performance can be sizable. Wise and DeMars (2005) synthesized a set of research studies focused on the relationship between motivation and test performance, finding that more motivated examinees tended to outperform their less motivated peers by an average of 0.58 standard deviations. Thus, obtaining a valid score requires both a well-developed test and an examinee who is willing to demonstrate what he knows and can do.

In a number of measurement settings, the issue of examinee motivation is usually not of great concern to test givers. It is reasonable in these contexts to assume that examinees are highly motivated to perform well because a high test score will help them attain something they desire, such as a course grade, graduation, scholarship, certification, or licensure. That is, examinees perceive meaningful personal consequences associated with their test performance. Moreover, if an unmotivated examinee chooses to behave non-effortfully, it is not viewed as the responsibility of the test giver, whose primary concern is developing and administering a valid test (i.e., one that is capable of obtaining a valid score).

There are many assessment programs, however, in which tests are administered to examinees who perceive few, if any, consequences associated with their performance. This type of *low-stakes* assessment (from the examinee's perspective) usually occurs in educational settings. For example, in the U.S. National Assessment of Educational Progress, students are

asked to take an assessment that has no bearing on their school grades, and for which they do not receive a score. In the absence of personal consequences, test-taking motivation is driven by internal examinee factors such as competitiveness or academic citizenship (Wise & Smith, 2011), which makes it difficult to assess how motivated a particular examinee is during a low-stakes test administration. Hence, in low-stakes testing contexts, motivation is far less certain and becomes a serious threat to the validity of test scores. Because of this validity threat, ensuring that examinees give their best effort becomes an additional responsibility of the test giver.

One of the unknowns with any low-stakes assessment is the percentage of unmotivated examinees. Estimation of this percentage is complicated, however, by the fact that there are varying degrees of motivation that an examinee might experience, and any classification scheme for identifying examinees whose test-taking behavior should be labeled "unmotivated" are necessarily based on arbitrary criteria. Nevertheless, if common criteria are applied across studies, some sense of the prevalence of unmotivated test-taking behavior might be gained.

A number of recent studies have used *response time effort (RTE*) index introduced by Wise and Kong (2005) as a criterion for classifying a test event as unmotivated. Studies using RTE, which is based on the proportion of rapid guessing behavior exhibited by examinees, have found considerable variation in unmotivated behavior. In K-12 settings, Wise, Kingsbury, Thomason, and Kong (2004) found 1% of students in grades 6-10 to be unmotivated when a .90 RTE value was used. In contrast, Wise, Ma, Kingsbury, and Hauser (2009), using a more stringent .85 RTE value, found the percentages of unmotivated examinees to be higher—ranging from 1% (in grade 3) to 7% (in grade 9).

In higher education settings, unmotivated test taking on low-stakes tests has generally been more prevalent than that found in K-12 settings. Based on an RTE value of .90, Wise and Kong (2005) identified 7% of their examinees as unmotivated, Wise and DeMars (2010) identified 11%, and Swerdzewski, Harmes, and Finney (2011) identified 26%. Using an RTE value of .87 for a series of assessments over the course of a university semester, DeMars (2007) found that the percentages of unmotivated examinees ranged from 1-25%. The sizable variation in the percentages of unmotivated students across these studies reflects the diverse impact of test characteristics, examinee characteristics, and contextual factors on the likelihood that a particular examinee will respond effortfully to a particular test.

Thus, examinee motivation represents a potentially serious threat to the validity of test scores. This problem is a distinctive aspect of educational measurement that complicates the science of measuring cognitive proficiency. Measurement in other fields, in contrast, rarely requires practitioners to be concerned about the motivation levels of their objects of measurement. As Cronbach (1960) noted over 50 years ago:

> In making a physical measurement—for instance, weighing a truckload of wheat—there is no problem of motivation. Even in weighing a person, when we put him on the scale we get a rather good measure no matter how he feels about the operation. But in a psychological test the subject must place himself on the scale, and unless he cares about the result he cannot be measured. (p. 52)

Effectively addressing the validity threat posed by examinee motivation involves pursuing two related goals. The first is to administer tests in a way that promotes examinee engagement and effort. The second is to detect instances in which examinees did not give good effort to their tests. These two goals are part of a general approach to low-stakes measurement

that recognizes that (a) we can enhance examinee motivation through our choices of the testing methods we adopt and (b) despite our efforts, there will always be some unmotivated examinees and we need to be able to identify them.

This paper is an integrative review of the role that a computerized adaptive test (CAT) can play in addressing these two goals. First, the issue of whether a CAT is more motivating is explored based on a review of relevant research. Next, the identification of non-effortful test taking is examined and a unique advantage of a CAT is illustrated using data from a widely used CAT program administered to U.S. school children. Finally, the promise of future generations of adaptive tests to enhance examinee motivation in low-stakes settings is discussed, drawing on the relevant research literature.

Are CATs More Motivating Than Conventional Tests?

It has been frequently claimed that because CATs administer items that are matched to an examinee's proficiency level, they are more motivating than conventional tests. The rationale for this claim is that lower proficiency examinees taking a CAT will not become frustrated or discouraged by items that are too difficult, while higher proficiency examinees will not be bored by items that are too easy. Thus, the exclusion of frustrating or boring items is expected to help maintain examinee motivation during a test event.

What is the evidence for this claim? The issue of whether a CAT is more motivating can be considered relative to two indicators. First, examinees might report that they perceived a CAT to be more motivating than a conventional test. Second, the relationship between motivation and test performance would suggest that if a CAT is more motivating, it should yield higher test performance relative to that yielded by a conventional test. While the second indicator is more stringent, if examinees reported higher motivation without accompanying

evidence of improved test performance, one might question the practical impact of higher reported motivation.

There is indirect support in the research literature on student motivation for the claim of increased motivation. This body of research has consistently found that the most intrinsically motivating tasks are those that individuals find moderately challenging (Pintrich & Schunk, 2002, Chapter 6). Additional support comes from research on a special type of computer-based test (CBT) call a *self-adapted test*. In a self-adapted test, before each item is administered, the examinee is allowed to choose its difficulty level from a set of discrete (usually 5-8) difficulty strata. Wise, Plake, Johnson, & Roos (1992) compared self-adapted tests and CATs, finding that examinees who were assigned to take a self-adapted test tended to choose difficulty levels that were similar to those they would have received on a CAT. That is, when they were given freedom to choose whichever difficulty levels they wished, most examinees chose levels whose items were about as difficult as those that would have been selected by a CAT algorithm. This finding suggests that the items administered by a CAT (i.e., those that are moderately challenging) are motivationally congruent with the difficulty levels that would be self-selected by examinees. In this way, the findings with self-adapted tests are consistent with research on intrinsic motivation.

Effects on Self-Reported Motivation

Several research studies have investigated the effects of CAT on self-reported motivation. Betz and Weiss (1976a) found that low-ability college-level examinees reported significantly higher levels of motivation on a stradaptive test (an early type of CAT) than on a paper-and-pencil test (PPT), while the reported motivation of high-ability examinees did not differ by test mode. They concluded that “the use of adaptive tests appears to result in comparable levels of

motivation in examinees differing in ability level" (p. 31). Pine, Church, Gialluca, and Weiss (1979) administered CATs and PPTs to a group of high-school students, finding higher mean levels of motivation reported on the CAT. Moreover, a stronger motivation effect was found for African-American students, leading the authors to conclude "it may be possible to obtain more comparable motivational states across racial groups using computer-administered tests" (p. 34). Thus, the two earliest studies suggest that a CAT may have the effect of creating a more homogeneous motivational environment across examinees, much as it yields proficiency estimates that are more homogeneous in precision. The only other study found on self-reported motivation (Arvey, Strickland, Drauden, & Martin, 1990) compared CAT and PPT versions of the Armed Services Vocational Aptitude Battery (ASVAB) that were administered to a sample of military recruits. They found higher levels of reported motivation on the CAT.

Thus, examinees across several different types of tests and examinee populations have reported that they found CAT to be more motivating. Nevertheless, one should be cautious in generalizing these findings. All three studies were conducted at a time when adaptive testing was a novel type of test, and it is likely that most, if not all, of the examinees were experiencing a CAT for the first time. This leads to questions about the degree to which the higher motivation could be attributable to a novelty effect, and the degree to which the motivational benefits would persist once examinees became more experienced with CAT remains unclear.

Effects on Test Performance

Additional evidence of a CAT motivation effect would be provided by empirical studies showing that test performance on a CAT was higher than to that yielded by a conventional test from the same measurement context (e.g., same measured construct, examinee population, etc.). The results from five meta-analyses comparing CBTs and PPTs, which are described below, are

informative to this question. Note, however, that although these meta-analyses focused on score comparability, their results provide information about the magnitude of a CAT motivation effect (which would actually represent evidence of *non-comparability*).

Because the results of the five meta-analyses were not completely consistent, it is helpful in comparing them to note that each examined a standardized mean difference effect size (ES) computed as $(\text{Mean}_{\text{CBT}} - \text{Mean}_{\text{PPT}})/\text{SD}_{\text{Pooled}}$. Positive ESs were consistent with a motivation effect, as they indicate that the higher mean occurred with the CBT. Negative ESs indicated that the PPT mean was higher.

Bergstrom's (1992) meta-analysis compared CATs and PPTs, synthesizing 20 ESs from eight research reports. The examinees in 12 of the ESs were adults, while the remaining 8 ESs were based on data from K-12 students. After deleting 5 ESs to attain a homogeneous set, Bergstrom reported a negligible mean ES of -.002.

Mead and Drasgow's (1993) meta-analysis examined 159 ESs comparing CBT and PPT performance on tests measuring the cognitive ability of young adults and adults. Of the 114 comparison studies they synthesized, 67 (59%) were CATs. They reported a mean ES for power tests of -0.03, which indicated that performance on the CBTs were slightly lower. Mead and Drasgow did not indicate whether adaptivity (i.e., whether the test was adaptive or non-adaptive) was a significant moderator of this mean difference ES. They did, however, report that adaptivity was not a significant moderator of the degree of correlation between CBTs and PPTs. Thus, the results from the first two meta-analyses provide little indication of a motivation benefit associated with a CAT.

The results from the other three meta-analyses suggest a more complicated story. Kim (1999) synthesized 226 ESs comparing CBTs and PPTs across a variety of samples and tests,

reporting a mean ES of +.019. Kim found that the set of effect sizes was heterogeneous, and that adaptivity was a significant moderator variable (indicating that there were significant differences in the CBT-PPT ESs for adaptive and non-adaptive CBTs). Specifically, CATs showed an ES of -0.15 while the ES for non-adaptive CBTs was +0.10, suggesting that adaptivity had a negative effect on test performance. Wang, Jiao, Young, Brooks, and Olson (2007) noted, however, that K-12 students represented only 4% of the samples in the Kim (1999) study. They conducted a meta-analysis focused on comparisons between CBTs and PPTs for mathematics tests administered to K-12 students. This analysis showed that, for the 36 ESs selected for analysis, the overall mean ES was -0.06. Adaptivity was found to be a significant moderator, however, with the ESs for non-adaptive CBTs and CATs being -0.09 and 0.08, respectively. This indicated that adaptivity had a positive effect. The Wang et al. meta-analysis was repeated for K-12 tests in reading (Wang, Jiao, Young, Brooks, & Olson, 2008), with similar findings. The overall ES in reading was -0.004, adaptivity was again found to be a significant moderator, and the respective ESs for non-adaptive CBTs and CATs were -0.01 and 0.05. Thus, the Wang et al. meta-analyses both provided evidence (albeit weak) consistent with the presence of a CAT motivation effect.

The mixed findings from five meta-analyses make it difficult to draw strong conclusions about the presence of motivational benefits from CATs. The two analyses focused on K-12 students provide the most promising evidence. It should be noted, however, that PPTs and CATs differ in multiple ways that might distort test performance. For example, CATs generally do not permit examinees to review (and possibly change) their answers, which could negatively affect performance. Additionally, there is some evidence that CATs can make examinees more anxious (Betz & Weiss, 1974a), which could also negatively affect performance. This suggests

that, relative to PPTs, CATs can affect examinees in multiple ways—some of which are performance enhancing and some of which are performance diminishing. Empirical comparisons of test performance between CATs and PPTs reflect the net effect of all of these influences, which makes it difficult to isolate the magnitude of a motivational effect. Nevertheless, because mean ESs found in the meta-analyses were consistently small in magnitude, it is probably safe to conclude that any motivational impact of a CAT on test performance is modest.

It is important to note that, whatever its magnitude, any potential motivational effect of a CAT on test performance is unlikely to be realized in operational use. This is because most CATs have been introduced in measurement contexts in which PPTs are already being used. Whenever both CAT and PPT versions of a test are administered, establishing the comparability of scores from the two versions is an important concern. For example, Guideline 22.1of the *International Guidelines on Computer-Based and Internet Delivered Testing* (International Test Commission, 2005) states that the scores from computerized and non-computerized versions should "produce comparable means and standard deviations or have been appropriately scaled to render comparable scores" (p. 11). In practice, this means that score comparability is often attained by equating the CAT scores to the scale established using the PPT. The implication of this is that, to the extent that a CAT increases motivation enough to improve test performance, *that improvement will be subsequently be equated away*. In essence, when CAT and PPT versions of a test are being used, the issue of whether the CAT possesses any motivational advantage is likely to be rendered moot by comparability concerns.

Conclusion

The claim that CATs are more motivating than conventional tests is frequently cited as an advantage of adaptive testing. The limited research on examinee self-reported motivation provides some support this claim. Moreover, the claim seems especially true for low-proficiency examinees, whose personal assessment histories tend to be characterized by test events in which the items were far too difficult for them. Hence, in this sense adaptive testing can be considered a more humane form of testing.

However, higher reported motivation has not conclusively been shown to be accompanied by higher test performance. There are several potential explanations for this. First, CATs may increase examinee motivation, but not be enough to significantly impact test performance. Second, any positive effects of motivation on test performance may be offset by other factors that have negative effects (such as not providing an opportunity to review/change answers). Third, it is probably the case that a CAT motivationally benefits only some examinees. In most measurement contexts, the percentage of examinees that are found to be unmotivated is less than 10 percent. In these situations, a CAT could markedly improve the test performance of those examinees without meaningfully increasing overall mean performance. Additional research targeted specifically at unmotivated examinees would yield more conclusive results about the extent to which increased motivation translates into improved test performance.

Can Unmotivated Test Taking be More Readily Detected on a CAT?

Because it is an internal state, an examinee's test-taking motivation cannot be directly measured. However, the examinee's motivational state influences the *effort* he directs toward his test, which is a behavior that can be measured. This suggests that the amount of effort expended during a test might be used as an indicator of an examinee's motivational state.

It should be noted that non-effortful test-taking behavior does not always indicate a lack of motivation. An examinee who is not feeling well might be highly motivated to do well on the test, while feeling too nauseated to give much effort. Alternatively, an otherwise motivated examinee may devote relatively little effort to a particular test item about a topic he has not yet had an opportunity to learn. Nevertheless, examinee effort generally provides a useful way to evaluate examinee motivation, and it can help identify instances when effort is so low that a score is not a trustworthy indicator of an examinee's level of proficiency.

Self-Report Measures

There are multiple ways to measure an examinee's test-taking effort. The most commonly used method is to ask the examinee, after the test has completed, to report his level of motivation or effort during the test. Short self-report instruments using a small number of Likert-style items, such as the Student Opinion Scale (Sundre & Moore, 2002) can produce reliable scores. The self-report method has several advantages. First, it is simple, economical, and can be administered in a short period of time (i.e., generally less than 2 minutes). Second, it can be used with both PPTs and CBTs. Because of its flexible ease of use, the self-report method has been the most widely used in research.

There are several disadvantages, however, to the self-report method. First, it is mildly intrusive, as examinees who have just completed their test are usually ready to focus on something other than the test they just took. Second, it is unclear how truthfully examinees will respond about their test-taking effort. Some examinees who did not give good effort might not want to admit it to test givers they respect, or if they fear punishment for reporting lack of effort. Alternatively, some examinees are predisposed to attribute failure to lack of effort, and they may falsely report low effort if they believed they did not do very well on their test. Third, self-report

measures provide a global assessment of effort, which makes it difficult to study any changes in effort that occur during a test event.

Measures Based on Test-Taking Behaviors

In recent years, there has been a growing interest in measuring effort based directly on behaviors exhibited by unmotivated examinees during a multiple-choice test event. There are several types of behaviors that might be observed. First, the examinee might simply omit items and not answer them. Second, he might give answers, but choose them randomly. Third, he might answer very rapidly in an attempt to get the test over with.

On a PPT, one can readily identify omitted items from an examinee's test booklet or answer sheet. In addition, one could compare his test performance to that expected by random responding. Unfortunately, omitted items and/or test performance resembling that from random responding might also be exhibited by low proficiency examinees, who would omit or guess at items because they did not know the correct answers. Thus, motivation and proficiency are potentially confounded when we try to interpret these types of behaviors with a PPT. In addition, it is virtually impossible with a PPT to identify instances in which an examinee answered an item very rapidly. Hence, when PPTs are used, the three test-taking behaviors that could indicate non-effortful test taking either provide ambiguous information or cannot be measured at all.

A CBT can yield information about both the number of omitted items and test performance. As with a PPT, however, a CBT is vulnerable to the same confounding between motivation level and proficiency level. This complicates interpretations of omitted items or low test performance as unambiguous indicators of low effort. It should be noted, though, that test

givers administering a CBT could prevent omitted items by requiring an examinee to answer each item before he can move on to the next.

In contrast to a PPT, a CBT can unobtrusively measure numerous behaviors during a test event. The most important of these is response time, defined as the time elapsed between when an item is displayed and when the examinee enters a response. Item response time, which can readily be measured and recorded, can be used to identify rapid guessing, which has been shown to be a relatively unambiguous indicator of test-taking effort (Wise & Kong, 2005). A rapid guess is one that occurs much faster than it should take an examinee to read, understand, and enter a response. Such responses get their name because of their characteristic accuracy rates that closely resemble those expected by chance under random responding.

A CAT, like all CBTs, can measure item response time, which can be used to identify rapid guessing, and permits assessment of effort down to individual item level. Omitted responses are typically not allowed on a CAT, because the item selection algorithm requires answers from all prior items in order to select subsequent items. A CAT, however, has an important advantage over a non-adaptive CBT. A distinctive feature of a CAT is that it can render low accuracy responses relatively unambiguous because it actively avoids administering items that are highly difficult for a given examinee. Assuming that the CAT (and its associated item pool) is able to administer items that are well targeted to an examinee's proficiency level, a set of responses with a very low accuracy rate can provide additional evidence of low effort.

To illustrate this, consider a hypothetical examinee taking a CAT whose items are calibrated using the Rasch model and each multiple-choice item has four response categories. Assuming that the CAT is well targeted, a motivated examinee should have a relatively consistent probability around .50 of passing items. An unmotivated examinee who guesses

randomly, in contrast, would have a .25 probability of passing items. The response of the examinee to a single item would not provide information sufficient to make a confident decision regarding whether the examinee was motivated or not. Over a set of items, however, one could gain confidence regarding whether the response pattern was more characteristic of a motivated examinee or an unmotivated one. For example, suppose that across a given set of ten items during a test event, the examinee passed two. Using the binomial theorem, the probability that a motivated examinee passed exactly two items (assuming a .50 probability on each item) would be .04. The probability that an unmotivated examinee passed exactly two items by random guessing would be .28. Hence, the outcome (two correct out of ten) is substantially more likely to have occurred if the examinee was unmotivated than if he was motivated. This is intuitively reasonable, because the observation that 20% of the items were passed seems much more likely to have come from an examinee who was randomly guessing than one who one who had a 50-50 chance of passing each item.

Thus, when a CAT is administered, test givers can potentially use the response times of individual items and response accuracy over sets of items to assess examinee effort. Table 1 summarizes the sources of effort-related information available under different types of tests. It is seen that CATs provide more information than CBTs, which in turn provide more information than PPTs, leading to the conclusion that non-effortful behavior can be most effectively identified when adaptive tests are used.

Effort Flagging of Test Events

How effort might be assessed on a CAT can be illustrated using the effort flagging criteria described by Wise, Ma, & Theaker (2012). They applied criteria being studied for use by Northwest Evaluation Association with its *Measures of Academic Progress (MAP*) adaptive

testing system. MAP is used to measure the academic growth of U.S. primary and secondary students in mathematics, reading, language arts, and science. MAP proficiency estimates are expressed as scale scores on a common scale that permits a student's growth to be assessed over time.

The five effort criteria are based on information from both item response time and response accuracy. If any of the effort flags were triggered for a test event, the score is classified as invalid due to low examinee effort. Two of the criteria are based on examinee behavior over the entire test event. Because examinees often exhibit non-effort during only a portion of a test event, however, three additional flagging criteria are based on subsets of the items during the test event.

Rapid guessing is identified based on the conceptualization that each item response can be classified as reflecting either *rapid-guessing behavior* or *solution behavior* (Schnipke & Scrams, 1997, 2002). This classification is done using pre-established time thresholds for each item using the normative threshold method (Wise & Ma, 2012) set at 10 percent. This means that the threshold for an item is set at 10 percent of the average time examinees have historically taken to answer the item. RTE for a test event equals the proportion of the examinee's responses that are solution behaviors (Wise & Kong, 2005).

The first two effort flagging criteria are based on RTE. The first uses the overall RTE for the test event:

<u>Flag A:</u> *If the examinee gave rapid guesses to at least 15% of the items (overall RTE ≤ .85).*

The second flag—designed to detect non-effort during only a portion of the test event—is based on rolling subsets of items. For example, for subsets of size 10, one would consider items

1-10, then 2-11, then 3-12, and so on, until the end of the test. Based on rolling subsets of size 10, an additional RTE-based flag was developed for evaluating low effort on a more local level:

Flag B: *If the examinee exhibited low RTE (local RTE ≤ .70) on at least 20% of the rolling subsets.*

The next two flagging criteria are based on response accuracy. Low-accuracy responses should be evaluated carefully to make sure that they were not due to the examinee receiving items that were much too difficult for him. It is important that the CAT item pool and selection algorithm be capable of administering items that are well targeted to an examinee's proficiency level throughout the test event. To accomplish this, a pool adequacy requirement is imposed specifying that low response accuracy will only be considered for CAT test events in which, at least 60% of the time, the examinee received an item whose difficulty was no more than three scale score points away from the examinee's momentary proficiency estimate[1]. This led to the development of two additional flags related to response accuracy:

Flag C: *If the examinee passed fewer than 30% of the items (overall accuracy ≤ .30) and at least 60% of all of the administered items were within three scale score points of the examinee's momentary proficiency estimate.*

Flag D: *If the examinee exhibited low accuracy (local accuracy ≤ .20) on at least 20% of the rolling subsets and at least 60% of all of the administered items were within three scale score points of the examinee's momentary proficiency estimate.*

The final effort flag is based on the joint occurrence of rapid guessing and low accuracy on any of the rolling subsets of items:

[1] Standard errors of examinee scores in reading are typically about 3.2 scale score points.

Flag E: *If the examinee passed no more than two items (local accuracy ≤ .20) and gave three or more rapid guesses (local RTE ≤ .70) on any 10-item subset, and at least 60% of all of the administered items were within three scale score points of the examinee's momentary proficiency estimate.*

The five flagging criteria were applied to a set of MAP test events. Testing records in reading from the fall and spring testing terms of the 2010-2011 academic year in a single U.S. state were retrieved from the Northwest Evaluation Association's *Growth Research Database*. The test records were limited to 287,690 students in grades 3-9 who were tested in both testing terms as part of their district sponsored testing programs.

The numbers of effort flags triggered during the fall test administrations are shown in Table 2. Over 31,000 (11%) of the test events were classified as invalid due to low effort. Nearly twice as many test events triggered rapid guessing flags as accuracy flags or joint rapid guessing and accuracy flags. The total number of flags triggered exceeded 47,000, indicating that it was common for multiple flags to be triggered by a test event.

Table 3 shows that when only a single type of flag was triggered, the type of flag varied. Rapid guessing flags occurred most often, followed by accuracy flags, with joint flags occurring relatively rarely. Particularly noteworthy is the finding that nearly a quarter of the test events triggering flags (7,622 out of 31,734) were identified only through accuracy flags. This underscores the unique advantage of CATs in identifying non-effortful behavior that is characterized by low accuracy even when rapid guessing does not occur.

Evidence for the validity of the flags can be found in the correlation of test scores with other variables if the scores from flagged test events are removed from the sample. The correlation between the entire sample's fall and spring RIT scores was .82. If examinees with flagged test events in either testing session are removed from the data, the correlation increased

to .86 even though the variances of both the fall and spring scores were reduced. These results are consistent with the idea that deleting the data from examinees exhibiting non-effortful behavior has the effect of improving the validity of a set of test scores by removing construct-irrelevant variance (Haladyna & Downing, 2004) from the data.

How Can CATs Be Modified to Better Address Examinee Motivation?

It is useful for test givers to have methods of detecting when examinees have exhibited low examinee effort, because they help identify scores with low individual score validity (Wise, Kingsbury, Hauser, & Ma, 2012). These methods, however, are designed to be applied *after* the test event has occurred. Consideration might also be given to what could be done during a test event to promote or maintain motivation. That is, to what extent can non-effortful behavior be reduced by the way the test event is conducted? The possibilities of what might be done range from relatively minor changes to the basic CAT algorithm and features to more fundamental changes in how we conceive of adaptive testing. Such changes require us to think beyond the traditional CAT algorithm and to consider testing methods that can have motivational benefits even in the absence of psychometric benefits.

One potential change in the CAT algorithm that has been investigated is the difficulty of the items that are administered to an examinee. Item selection in a CAT is typically driven largely by an item response theory-based maximum information criterion, which is intended to maximize the efficiency of test events and yield scores with maximal precision. This results in test events in which examinees pass their items about 50% of the time. Some researchers have argued that this success rate is too low and threatens the motivation of examinees who are used to passing a much higher percentage of items when they take tests.

Studies of CATs targeted at higher success rates (usually 60%, 70%, and 80%) have found that examinees do report higher levels of motivation compared to that from a CAT using a 50% success rate, but without significant motivational benefits in terms of higher test performance (Bergstrom, Lunz, & Gershon, 1992; Häusler & Sommer, 2008; Lunz & Bergstrom, 1994; Tonidandel, Quiñones, & Adams, 2002). In addition, these studies have demonstrated that CATs using higher success rates yield proficiency estimates with higher standard errors than those from traditional 50% CATs. Thus, there is a tradeoff between motivation and precision; examinees tend to find higher success rates somewhat more motivating, but at the cost of slightly less precise proficiency estimates. Because higher motivation has not been found to result in higher performance, however, precision has been a priority for test givers and few, if any, operational CATs use a higher success rate. Nevertheless, despite arguments that measurement efficiency is the primary reason for a CAT (Wainer, 1993), success rate should be recognized a factor that could be potentially be manipulated to enhance examinee motivation.

An additional feature that could potentially be provided to examinees taking a CBT is item feedback regarding the correctness of their responses. The rationale for providing this feature is that, by providing item feedback, examinees' engagement will be maintained because they will be motivated to see if they answered items correctly or not. On a high-stakes fixed-length test, providing item feedback would naturally raise questions about the validity threat posed by exposing item content. But with a CAT (using large item pools) in a low-stakes testing context (in which test-taking motivation is apt to be of greatest concern), providing item feedback would likely raise fewer item exposure concerns.

The research findings on the relationship between item feedback and motivation are mixed. Betz and Weiss (1976a) found that low-ability examinees reported lower motivation

when item feedback was provided, while high-ability examinees reported higher motivation. Similarly, Betz and Weiss (1976b) found that item feedback resulted in higher test performance for the high-ability group, but lower performance for the low-ability group. Pine et al. (1979) found that African-American examinees reported more negative attitudes toward item review and indicated that item review made them nervous and interfered with their concentration. These results suggest that the impact of item feedback varies across examinees such that some may find it motivating while others may find it anxiety producing or distracting. The mixed impact of item review on motivation is similar to the mixed impact of item feedback on test performance [see Vispoel (1998) for a review of this research]. Thus, test givers trying to improve motivation should use item feedback cautiously, because it may induce other types of construct-irrelevant factors such as anxiety.

Expanded Types of Adaptive Tests

Standardization is one of the foundational concepts of modern measurement. It embodies the "principle that valid interpretation of test scores relies on the expectation that every test administration has been conducted under the same, standardized conditions of measurement" (McCallin, 2006, p. 634). The primary goal of standardization is to reduce the impact of group-level construct-irrelevant variance on test scores. For example, use of a standard time limit with a test ensures that differential time limits do not introduce construct-irrelevant variance that can compromise interpretations of test scores.

Some construct-irrelevant factors, however, do not affect all of the examinees in a group. Person-specific construct-irrelevant variance is systematic error that can affect an individual examinee's score (Haladyna & Downing, 2004). Test-taking motivation is an example of

person-specific construct-irrelevant variance that negatively affects individual examinee scores, and thus threatens individual score validity.

Person-specific construct-irrelevant variance due to internal factors such as motivation will generally not be reduced through test standardization. Instead, it might be effectively managed through an *individualization* of the test administration. Specifically, a test might adapt to the presence of a construct-irrelevant factor during a particular examinee's test event by altering the test administration in a way that might mitigate the effect of the validity threat. Cronbach (1960) considered the management of construct-irrelevant factors a particular type of standardization:

> We may better understand the problem of framing directions and arousing motivation if we realize that the psychometric tester tries to standardize the behavior of the subject, as well as the test stimuli. Even though he is measuring individual differences, his procedures are designed to *eliminate* individual differences—to eliminate, that is, variation in every characteristic save the one that his test is supposed to measure. (p. 59-60)

A good example of individualization relevant to motivation is the *effort-monitoring CBT* introduced by Wise, Bhola, and Yang (2006). In this type of test, the computer algorithm monitors item response times. If it detects that an examinee has begun to exhibit rapid-guessing behavior, a message is displayed to the examinee noting that a decrease in effort has been detected and encouraging the examinee to increase his or her effort. Wise et al. found that those examinees receiving messages exhibited increased levels of motivation as indicated by longer response times and higher success rates on subsequent items). A replication of the Wise et al. study by Kong, Wise, Harmes, & Yang (2006) found that examinees receiving effort messages

performed significantly better on subsequent items than examinees who deserved, but did not receive messages.

In an effort-monitoring CBT, messages are displayed only to those whose test-taking behavior warrants them. The remaining examinees do not receive messages. In this way, the effort-monitoring CBT is less standardized than a conventional PPT or CBT because its intervention is conditional on the behavior of the examinee[2]. It focuses on maximizing the validity of individual test scores by identifying and reducing construct-irrelevant factors that affect particular examinees rather than maintaining a standard test administration for all examinees. Such a strategy is designed to maximize the collective individual score validity of the scores from a group of examinees.

In what other ways might a CAT adapt to unmotivated examinees other than through manipulation of item difficulty? One answer is that future CATs might select or present items based on a pre-knowledge of an examinee's interests (which would presumably be more motivating). This would require items that could be framed within a variety of contexts. If it were known in advance that a particular examinee liked fishing, for example, a particular mathematics item might be framed in the context of a fishing example or scenario. If a different examinee liked sewing, the same item might be presented in a sewing context. The challenge to the test giver would be to develop items that could be plausibly re-framed according to the examinee's interests (the motivational need) while maintaining reasonably consistent difficulty levels (the psychometric need). This idea is speculative, however, and research is needed to understand its feasibility.

---

[2] A traditional CAT can also be considered individualized relative to a conventional test in that the difficulty levels of the items administered depend on the correctness of an examinee's responses to earlier items. The result is that each examinee receives a test whose item difficulty levels are adapted to his or her level of proficiency.

Conclusions

Wainer (1993) cautioned that "if we are to improve the practice of testing, we must allow later generations of tests to be better than earlier ones" (p. 19). CATs represent a major advancement in how efficiently we can gather information about examinee proficiency. While testing efficiency has been a primary reason for implementing CAT programs in education, there is evidence that the targeting of item difficulty to examinee proficiency brings with it modest motivational benefits that can improve individual score validity. It is important, however, that these motivational benefits not be equated away when CATs are used in conjunction with conventional tests (i.e., PPTs or non-adaptive CBTs). Understanding that CATs can improve score validity by reducing construct-irrelevant variance in ways that conventional tests cannot may lead to recognition that test performance differences between testing modes might be viewed as a positive result rather than a threat to comparability that must be corrected for.

Unmotivated examinees represent an individual score validity threat that is likely to be present whenever low-stakes tests are used in education. Without personal consequences associated with test performance, not all examinees will be motivated to give their best effort. It is therefore important that we have effective methods for identifying instances of low examinee effort, because they will indicate test events whose scores have low individual score validity. It has been shown in this paper that CATs have clear advantages over conventional tests in identifying such instances—a finding that enhances the value of CATs to those administering low-stakes tests.

A CAT exemplifies the psychometric advantages that can be obtained when a test is less standardized than its conventional counterpart. Additional validity-related advantages might be found in a new generation of CATs that can adapt to the motivational/affective states of

examinees. Developments in this direction will require a broader perspective on what it means for a test to adapt. It will also require a consideration of the potential validity gains that might be obtained by adopting individualized test practices that are directed toward reducing the impact of construct-irrelevant variance on test scores.

In low-stakes testing programs, test givers have the dual responsibilities of providing a high-quality test that is capable to yielding valid scores and motivating examinees to give good effort to that test. This paper has shown that, by its unique nature, a CAT encourages examinees to maintain engagement during a test while allowing us to better detect instances when disengagement has occurred.

References

Arvey, R. D., Strickland, W., Drauden, G., & Martin, C. (1990). Motivational components of test taking. *Personnel Psychology, 43*, 695-716.

Bergstrom, B. (1992, April). *Ability measure equivalence of computer adaptive and and paper and pencil tests: A research synthesis.* Paper presented at the annual meeting of the American Education Research Association, San Francisco.

Bergstrom, B. A., Lunz, M. E., & Gershon, R. C. (1992). Altering the difficulty in computer adaptive testing. *Applied Measurement in Education, 5*, 137-149.

Betz, N. E., & Weiss, D. J. (1976a). *Psychological effects of immediate knowledge of results and adaptive ability testing* (Research Report 76-4). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Betz, N. E., & Weiss, D. J. (1976b). *Effects of immediate knowledge of results and adaptive ability testing on ability test performance.* (Research Report 76-3). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Cronbach, L. J. (1960). *Essentials of psychological testing* (2nd Ed.) New York: Harper & Row.

DeMars, C. E. (2007). Changes in rapid-guessing behavior over a series of assessments. *Educational Assessment, 12*, 23-45.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23(1)*, 17-27.

Hauser, C., Kingsbury, G, G., & Wise, S. L. (2008, March). *Individual validity: Adding a missing link.* Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

Häusler, J., & Sommer, M. (2008). The effect of success probability on test economy and self-confidence in computerized adaptive tests. *Psychology Science Quarterly, 50*, 75-87.

International Test Commission (ITC) (2005). *International guidelines on computer-based and internet delivered testing* (Version 2005). Available from http://www.intestcom.org

Kim, J. (1999, October). *Meta-analysis of equivalence of computerized and P&P tests on ability measures.* Paper presented at the annual meeting of the Mid-Western Educational Research Association, Chicago.

Kingsbury, G. G., & Hauser, C. (2007, April). *Individual validity in the context of an adaptive test.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Kong, X. J., Wise, S. L., Harmes, J. C., & Yang, S. (2006, April). *Motivational effects of praise in response-time based feedback: A follow-up study of the effort-monitoring CBT.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Lunz, M. E., & Bergstrom, B. A. (1994). An empirical study of computerized adaptive testing conditions. *Journal of Educational Measurement, 31*, 251-263.

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin, 9*, 287-304.

McCallin, R. C. (2006). Test administration. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development*. Mahwah, NJ: Lawrence Erlbaum Associates.

Pine, S. M., Church, A. T., Gialluca, K. A., & Weiss, D. J. (1979). *Effects of computerized adaptive testing on black and white students.* (Research Report 79-2). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Pintrich, P. R., & Schunk, D. H. (2002). *Motivation in education: Theory, research, and applications* (2nd ed.). Upper Saddle River, NJ: Merrill Prentice-Hall.

Schnipke, D. L. & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Education Measurement, 34(3),* 213-232.

Schnipke, D.L. & Scrams, D.J. (2002). Exploring issues of examinee behavior: Insights gained from response-time analyses. In C. N. Mills, M. T. Potenza, J. J. Fremer, & W. C. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments.* Mahwah, NJ: Lawrence Erlbaum Associates.

Sundre, D. L., & Moore, D. L. (2002). The Student Opinion Scale: A measure of student motivation. *Assessment Update, 14(1),* 8-9.

Swerdzewwski, P. J., Harmes, J. C., & Finney, S. J. (2011). Two approaches for identifying low-motivated students in a low-stakes assessment context. *Applied Measurement in Education, 24*, 162-188.

Tonidandel, S., Quiñones, M. A., & Adams, A. A. (2002). Computer-adaptive testing: The impact of test characteristics on perceived performance and test takers' reactions. *Journal of Applied Psychology, 87*, 320-332.

Vispoel, W. P. (1998). Psychometric characteristics of computer-adaptive and self-adaptive vocabulary tests: The role of answer feedback and test anxiety. *Journal of Educational Measurement, 35*, 155-167.

Wainer, H. (1993). Some practical considerations when converting a linearly administered test to an adaptive format. *Educational and Psychological Measurement, 12(1),* 15-20.

Wang, S., Jiao, H., Young, M. J., Brooks, T. E., & Olson, J. (2007). A meta-analysis of testing mode effects in Grade K-12 mathematics tests. *Educational and Psychological Measurement, 67*, 219-238.

Wang, S., Jiao, H., Young, M. J., Brooks, T. E., & Olson, J. (2007). Comparability of computer-based and paper-and-pencil testing in K-12 reading assessments: A meta-analysis of testing mode effects. *Educational and Psychological Measurement, 68*, 5-24.

Wang, T. & Kolen, M. J. (2001). Evaluating comparability in computerized adaptive testing: Issues, criteria, and an example. *Journal of Educational Measurement, 38,* 19-49.

Wise, S. L., Bhola, D., & Yang, S. (2006). Taking the time to improve the validity of low-stakes tests: The effort-monitoring CBT. *Educational Measurement: Issues and Practice 25(2),* 21-30.

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10,* 1-17.

Wise, S. L., & DeMars, C. E. (2010). Examinee non-effort and the validity of program assessment results. *Educational Assessment, 15,* 27-41.

Wise, S. L., Kingsbury, G. G., Hauser, C., & Ma, L. (2012). *How do I know that this score is valid? The case for assessing individual score validity.* Manuscript submitted for publication.

Wise, S. L., Kingsbury, G. G., Thomason, J., & Kong, X. (2004, April). *An investigation of motivation filtering in a statewide achievement testing program*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego.

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18,* 163-183.

Wise, S. L., & Ma, L. (2012, April). *Setting response time thresholds for a CAT item pool: The normative threshold method.* Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, Canada.

Wise, S. L., Ma, L., Kingsbury, G. G., & Hauser, C. (2010, May). *An investigation of the relationship between time of testing and test-taking effort.* Paper presented at the annual meeting of the National Council on Measurement in Education, Denver, CO.

Wise, S. L., Ma, L., & Theaker, R. A. (2012, May). *Identifying non-effortful student behavior on adaptive tests: Implications for test fraud detection.* Paper presented at the Conference on the Statistical Detection of Potential Test Fraud, Lawrence, Kansas.

Wise, S. L., Plake, B. S., Johnson, P. L., & Roos, L. L. (1992). A comparison of self-adapted and computerized adaptive tests. *Journal of Educational Measurement*, 29, 329-339.

Wise, S. L., & Smith, L. F. (2011). A model of examinee test-taking effort. In J. A. Bovaird, K. F. Geisinger, and C. W. Buckendal (Eds.) *High-stakes testing in education: Science and practice in K-12 settings (pp. 139-153).* Washington, DC: American Psychological Association.

Table 1

Availability of Indicators of Effort Under Different Test Types

	Test Type		
Effort Indicator	PPT	Non-Adaptive CBT	CAT
Self-report	Yes	Yes	Yes
Response Time	No	Yes	Yes
Response Accuracy	No	No	Yes

Table 2

Numbers of Test Events with Different Types of Effort Flags Triggered

Effort Flag Type	N	% of Total Sample
Zero Flags	255,956	89.0
At Least One Flag	31,734	11.0
Rapid Guessing Flag (Overall or Local)	22,398	7.8
Accuracy Flag (Overall or Local)	11,730	4.1
Joint Rapid Guessing and Accuracy Flag	13,330	4.6

Table 3

Numbers of Test Events in Which Only A Single Type of Flag was Triggered

Effort Flag Type	N	% of Total Sample
Rapid Guessing Flag (Overall or Local)	10,584	3.7
Accuracy Flag (Overall or Local)	7,622	2.6
Joint Rapid Guessing and Accuracy Flag	1,361	0.5

How Do I Know That *This* Score is Valid? The Case For Assessing Individual Score Validity

Steven L. Wise, G. Gage Kingsbury, Carl Hauser, and Lingling Ma

Northwest Evaluation Association

Author Note

Correspondence concerning this article should be addressed to Steven L. Wise, Northwest Evaluation Association, 121 NW Everett St., Portland, OR 97209. E-mail: steve.wise@nwea.org

Abstract

Whenever the purpose of measurement is to inform an inference about an individual student, it is important that we be able to trust that the student's test score accurately reflects what that student knows and can do. Such trust comes from the dual requirements that (a) we are using a competently developed test that is capable of providing valid scores from the target population of students and (b) that the student's test event was not influenced by construct-irrelevant factors that could meaningfully distort his or her score. Although traditional validity evidence is useful for informing the viability of the first requirement, it provides little information regarding the second. In this paper we propose a process for collecting evidence of individual score validity, which complements traditional validity evidence in assessing the trustworthiness of individual test scores.

How Do I Know That *This* Score is Valid? The Case For Assessing Individual Score Validity

Whenever we administer a test to measure a student's level of proficiency in some educational domain, our primary goal is to obtain a valid score for that student. That is, we seek information that we can use to help us make inferences regarding what that particular student knows and can do. Because of the pervasive and central role that test scores play in a variety of educational inferences about students (e.g., academic progress, graduation, entrance into higher education institutions), it is important that we have methods for evaluating the trustworthiness of such information. This implies that whenever our unit of inference is the individual student, any inquiry regarding the trustworthiness of test data must at some point focus on the individual student scores.

This idea of trustworthiness of information from test scores suggests the fundamental measurement concept of test score validity. Although test score validity has engendered much debate and research for many decades (Cronbach, 1971; Kane, 2006; Messick, 1989), there is general consensus regarding the types of evidence that are typically considered a valuable part of a validity argument. It is our view, however, that although the types of evidence that are traditionally collected to validate score-based inferences are highly important, they are not sufficient to answer the question of whether a particular test score is valid. Our purposes in this paper are to (a) examine the limitations of traditional validation methods, (b) propose a practical framework for collecting additional evidence to support the validation of individual test scores, (c) provide examples of the proposed validation framework, and (d) discuss additional directions for research.

To illustrate our perspective, consider two different testing scenarios. The first is an instance of an individually administered achievement or aptitude test, such as that routinely administered by a school psychologist. In this scenario, suppose that the test administrator is using a test for which a substantial amount of traditional validity evidence has been accrued. The set of items on the test is well aligned to the content domain that the test was designed to measure and, collectively, the items provide a good representation of the content domain. In addition, previous research has supported the internal structure of the set of test items, and scores from the test have been found to demonstrate good convergent and discriminant validity evidence.

But despite the varied and substantial amount of evidence that is available supporting the validity of the test's use, when the school psychologist is testing a particular student, she attends closely to the student's behavior. She has been trained to continually monitor the test administration and either terminate the test or take some sort of corrective action if she feels that the testing session is unlikely to yield a test score that validly represents what the student knows and can do. A variety of circumstances might trigger such a response. The student might appear disengaged and unmotivated during the test, or might appear fatigued or ill. Being tested might make the student so anxious that test performance is likely to be adversely affected. The testing room might unexpectedly become very warm or noisy during the session, or there might be the interruption of a fire drill that disrupts the student's concentration. The administrator might realize during the test session that some type of testing accommodation should have been provided to this student, but had been overlooked. In essence, the test administrator has a professional responsibility to look for evidence that

that student's test performance is not a trustworthy indicator of the student's proficiency level, and to consider invalidating that score if the evidence warrants it.

There are two related points to note regarding this scenario. First, the events and behaviors that the test administrator is watching for are all *construct-irrelevant* factors (Messick, 1984) that threaten the validity of test score interpretations. The presence and impact of these construct-irrelevant factors commonly varies across students, which has the effect of increasing test score variance. Haladyna and Downing (2004) generically referred to these types of validity threats as construct-irrelevant variance (CIV), and provide an excellent discussion of the various forms that they might take in high-stakes testing. In the present scenario, however, when monitoring the test session the administrator is focused primarily on student-related CIV. Second, traditional validity research is of limited use in answering the specific question of whether or not a particular test score is valid. For example, although research may have indicated that being tested can produce high levels of test anxiety in some students, the observation and judgment of the test administrator is needed to determine whether test anxiety is threatening the validity of the score from *this* particular student during *this* particular test event.

Contrast the individual testing scenario with the more familiar scenario of group-administered achievement testing, in which a common test is simultaneously administered to group of students. Group testing emerged nearly a century ago out of the need to efficiently assess the mental abilities of a large number of military recruits during World War I (Anastasi, 1988). Soon thereafter, this form of testing became the dominant method for educational testing in the United States.

An important aspect of group-administered testing is that only a limited number of proctors are typically used. Instead of the one-to-one relationship between the administrator and the student that characterizes individual testing, there is a one-to-many relationship in group testing, which constrains the extent to which proctors can closely monitor the test-taking behavior of individual students. Because it is unrealistic to expect proctors to closely monitor the test events for all individual students, proctor responsibilities typically do not extend beyond distributing test materials, ensuring that the test is administered in the prescribed standardized manner, and deterring and/or detecting cheating behavior by students.

Thus, there is a key distinction between the two types of test events. Under individual testing, the person giving the test has a responsibility to monitor the test session and invalidate test scores if there is sufficient evidence of CIV. In contrast, under group testing, test proctors are generally not assigned (and, more importantly, could not realistically fulfill) such responsibility. This reduced capacity to assess the validity of individual test scores represents a crucial validity-relevant cost associated with the advent of group testing.

The Limitations of Traditional Validity Evidence

A number of definitions of validity have been proposed, but they generally refer to test scores in the plural, rather than singular, sense. For example, the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) defines validity as, “the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests” (p. 9). The validity of individual test scores is generally not emphasized in these definitions.

Consistent with a chosen validity definition, measurement practitioners seek to develop an argument for score validity. The evidence underlying this argument can come in a variety of forms. First, evidence can be collected relevant to the content of the test items. This type of evidence typically consists of expert judgments about the degree to which a set of items represents the target construct or is aligned with the content domain. Second, whenever there is an identifiable (and measurable) criterion variable that test scores are expected to predict, evidence can be collected regarding the correlation between test scores and that criterion. A third type of evidence concerns the theory-based expectations regarding the test scores. This type of evidence includes (a) examinations of the internal structure of the test, (b) the pattern of correlations between the test scores and measures of other constructs, and (c) the degree to which groups of students that theoretically should differ on the target construct do, in fact, exhibit mean score differences.

What is deficient about this traditional approach to test score validation? Its primary limitation is that the strategies used to collect validity evidence typically provide little, if any, information concerning validity threats posed by the presence of student-related CIV. To illustrate this, consider a group of students taking a standardized achievement test in math. If a particular student is not motivated to do his best, and consequently devotes little effort toward the test, the resulting test score would be invalid if it markedly underestimates his actual level of math proficiency. Could traditional validity evidence be useful in identifying this source of score invalidity? Content-related evidence would not be helpful because it looks only at the item content, and not at how students behave when administered the items. Criterion-related evidence would not be sensitive to the presence of unmotivated students unless there was something about the test that routinely elicited low effort from enough

students that the correlation was substantially affected. In such a case, the correlation between the criterion and test score would be low, which would suggest that the test could not be trusted to yield valid scores. Note, however, that (a) the correlation would be of virtually no use in identifying exactly which students exhibited did not good effort and (b) the correlation could be quite high if the proportion of unmotivated students was low. Hence, at best, correlations provide limited guidance to us when construct-irrelevant factors affect some students, but not others. Similarly, construct-related validity evidence, which relies on correlations and mean difference effect sizes, would also be of limited use in identifying infrequent instances of score invalidity due to CIV.

It should be emphasized that we are not claiming that traditional validity evidence is not useful. On the contrary, traditional evidence is essential for any argument supporting general inferences made regarding the scores from a test. Rather, we are asserting that traditional evidence, as important as it is, is insufficient to support the validity of individual test scores. This underscores the need for additional validity information that is specifically focused on individual test events. In essence, we are proposing that methods be explicitly employed for detecting the presence of "alternative explanations" (Mislevy, 2009, p. 97) that can weaken the inferences in an assessment use argument.

Individual Score Validity

When we administer a test, we have only a limited amount of control over the factors that affect score validity. This idea might be best understood if a test event is conceptualized as a series of encounters between a *student* and a *test item* within a particular testing *context*. Although we have a great deal of control over the presentation of the test item, we have far less control over the student, who might be affected by a variety of factors such as lack of

motivation, test anxiety, illness, fatigue, or a willingness to cheat to attain a high score. Moreover, even though we try to standardize test administration as much as possible, in practice the testing context is usually only partially under our control because there is inevitably some variability in the testing context that can exacerbate the impact of student-related factors. For example, students tested just before lunch might be less motivated than those tested at other times of the day, a particularly stern-looking test proctor might elicit greater degrees of test anxiety than other proctors, or the layout of a particular testing room might be more tempting than other room layouts to students inclined to cheat.

In this paper, we define *individual score validity (ISV)* as the degree to which an individual test score is free from sources of CIV. As such, it serves as a complement to traditional validity evidence. That is, traditional validity evidence pertains to the adequacy of a test's items for a specified population of students and the general conditions under which the items are administered and scored. ISV focuses on additional factors that may threaten score validity, even when there is clearly adequate traditional validity evidence. These factors pertain predominantly to the less controllable student-related factors described above.

The current work builds on previous research on the validity of individual scores. Kingsbury and Hauser (2007) introduced the term *individual validity*, which they generally referred to as, "the extent to which an individual's score has utility" (p. 1). Later, Hauser, Kingsbury, and Wise (2008) defined individual validity as, "the process of identifying the extent to which a test score for an individual was a reasonable indicator of the individual's current ability on the content of interest" (p. 1-2). In this paper, we have slightly amended the terminology to emphasize that the validity inference concerns the individual's score, and not the individual himself.

A Process for Investigating ISV

Due to the myriad potential sources of CIV, it is practically impossible to identify with certainty which scores have been meaningfully affected by construct-irrelevant factors and which ones have not. In a particular measurement context, however, we can usually make reasonable judgments regarding which construct-irrelevant factors pose the greatest ISV threat, and then attempt to identify the test events that have been meaningfully affected by those factors. For example, in a low-stakes achievement testing context, low student motivation would typically pose a major ISV threat, while cheating would not. In contrast, in the context of a high-stakes licensure test in a medical field, cheating or test-taker anxiety would pose far more serious ISV threats than would low motivation. Thus, the process of seeking evidence for ISV becomes one of investigating whether one or more pre-identified sources of CIV influenced the test event under study. This process can be described in six steps.

Step 1: Identify the Target ISV Threats

Which sources of CIV are most likely to pose threats to ISV for a particular test event? The taxonomy of CIV sources provided by Haladyna and Downing (2004) should provide a useful guide, from which we should be able to identify the sources of greatest concern. The outcome of this step defines the focus of the remaining steps.

Step 2: Develop Indicators for Each ISV Threat

Identifying the presence of a particular source of CIV in a given test event represents the largest challenge in evaluating ISV. To that end, we must be able to identify suitable indicators of the CIV source. In addition, if individual test scores will be classified as invalid on the basis of these indicators, it is important that validity evidence can be provided for use

of the indicators. Because classifying a score as invalid is a nontrivial matter, even in low-stakes assessment, we must be able to justify the reason for the ISV decision.

Step 3: Establish Criteria for Classifying a Score as Invalid

After indicators of CIV have been chosen, we should also establish criteria for classifying a test event as invalid. For example, if we are using a self-report measure of test anxiety to identify those test events that were unduly influenced by student anxiety (and therefore were likely to negatively bias the student's test score), we should operationally define procedural rules for invalidating scores (e.g., if the mean anxiety score is greater than some value *x*, then the score will be deemed invalid).

Step 4: Analyze the Test Events and Flag Invalid Events.

After the CIV indicators have been chosen and criteria established for invalidating test scores, the next step is to analyze a set of test events to identify which test scores (if any) have been flagged (i.e., classified) as invalid.

Step 5: Take Corrective Action Regarding Invalid Test Events

Once invalid test scores have been identified, an additional step is to decide what to do about them. This generally involves policy decisions, for which there are often a range of options. For example, we might choose to not release the scores at all, re-test the students, indicate on score reports which scores are untrustworthy, statistically adjust the test scores to remove the effects of the CIV, or impose some type of sanctions on students (e.g., in the case of cheating).

Step 6. Evaluate the validity of the method for classifying test events as indicating low ISV.

In this step, the validity of the flagging process is considered. Is there evidence to support a conclusion that the flagged test events have been meaningfully affected by the CIV source under consideration? For example, if the flagged test events are removed, do the remaining events exhibit stronger traditional validity evidence?

Example 1: An Adaptive Test of Reading Proficiency

The context of this example is an interim testing program in reading. Testing records in reading from the fall and spring testing terms of the 2010-2011 academic year in a single U.S. state were retrieved from the Northwest Evaluation Association's *Growth Research Database* (GRD). The test records were limited to 287,690 students in grades 3-9 who were tested in reading during both fall and spring as part of their district sponsored testing programs. All tests were part of NWEA's *Measures of Academic Progres*s (MAP) testing system, which administers computerized adaptive tests (CATs). MAP proficiency estimates are expressed as scale scores (RIT) on a common scale that permits a student's growth to be assessed when he is tested at different times.

ISV Threat: Student Test-Taking Effort

As the first step of the ISV process, we identified an ISV threat that was of particular concern with the MAP assessment. Obtaining a valid test score from a student involves more than just administering a high-quality test under standardized conditions. A valid score also requires a motivated student who behaves effortfully throughout his test event. The fact that this is not always the case represents an ISV threat. Because item responses provide the basis for estimating the student's proficiency, any responses that do not reflect the student's

proficiency level will distort proficiency estimation, and thereby diminish the validity of the student's score. In addition, because non-effortful responses tend to be incorrect more often than those for which the student exhibits good effort, the presence of non-effortful behavior will tend to result in a negatively biased proficiency estimate.

In MAP, a student's growth is measured using his RIT score at Time 2 minus his RIT score at Time 1. In practice, the observed growth scores are consistent with the nature and distribution of growth that we would expect to see in these students. However, we sometimes observe growth score values that are either unreasonably low or unreasonably high. For students whose true growth between the two test events is low relative to the size of the measurement errors inherent in their RIT scores, the instability in the RIT scores at each time point will sometimes result in negative growth scores being observed by chance. In contrast, we occasionally see negative growth scores whose magnitude exceeds what we could reasonably attribute to low growth and the presence of measurement error in each of the test scores. Similarly, we sometimes observe positive growth scores whose magnitude far exceeds the amount of actual academic growth that a student could realistically make during that time period. Both forms of extreme growth—but especially the negative form—are difficult for teachers, parents, and students to understand, and their occurrence reduces the teacher's ability to make useful instructional decisions concerning the students in question.

Extreme growth scores have been shown to be associated with low student effort on one of the two testing occasions whose scores are used to construct a growth score (Wise & Ma, 2012). Specifically, low effort at Time 1 would introduce a construct-irrelevant factor that could negatively bias test performance. If the student then gave good effort at Time 2,

then one would expect to see a growth score that overestimates the amount of academic growth the student had actually attained. And because low test-taking effort can have a sizable effect on test performance (Wise & DeMars, 2004; 2010), this overestimated growth can be substantial in magnitude. Likewise, if a student gave good effort at Time 1, but low effort at Time 2, one could expect to see sizable negative growth. Therefore, at Step 1 of the ISV process, student effort is considered a key CIV source that threatens the validity of individual growth scores.

The second step of the process was to identify indicators of the ISV threat. In choosing these indicators, we focused on two characteristic behaviors of students exhibiting low test-taking effort. First, they often respond very rapidly to items—more rapidly than it would take to even read the item (Wise & Kong, 2005). This type of response behavior has been termed *rapid-guessing behavior*, with all other responses being termed *solution behavior* (Schnipke & Scrams, 1997; 2002). Wise and Kong (2005) developed an index of test-taking effort termed *response time effort (RTE)*. Second, non-effortful student responses tend to exhibit accuracy rates similar to those from random responding (Wise & Kong, 2005). Both RTE and response accuracy provide the bases of the effort indicators.

The third step of the ISV process was to establish criteria for identifying invalid test events. We developed three effort criteria for use in flagging student test events as indicating non-effort. These criteria, the details of which are found in the Appendix, are based on RTE and response accuracy. In our effort analysis of CAT data, a student's test event was classified as invalid (i.e., as untrustworthy) if any of the three flags had been triggered. In addition, the total number of flags triggered by a given test event was studied as a potential indicator of the degree of non-effort that occurred.

The fourth step in the ISV process was to evaluate the MAP reading data and identify test events whose scores were classified as invalid due to low effort. Table 1 shows the rates of occurrence for the various effort flags at each time period. There were several noteworthy aspects to these results. First, more flags were triggered in the fall than in the spring. Second, many of the test events that triggered any of the flags triggered multiple flags, as 23,952 students triggered at least one flag in the fall, while the total number of flags triggered exceeded 51,000. Similarly, in the spring, 16,423 of the test events triggered flags while the total number of flags triggered exceeded 34,000. Finally, the flags based solely on response time (flags A & B) were triggered more often than the flag based on the joint occurrence of rapid guessing and low response accuracy (flag C).

Having identified students who triggered effort flags, the fifth step of the ISV process was to take corrective action. Two types of policy decisions were made. The first was a decision to include ISV information on MAP score reports by indicating scores classified as invalid. The second was a decision to exclude from item calibration analyses and norms studies the data from invalid test events.

The sixth step of the ISV process was to evaluate the validity of the flagging process itself. This was done by assessing the impact of removing invalid scores from the sample. Table 2, which focused on students with invalid scores in the fall only, shows the impact of removing their scores from the sample. The removal of students who triggered flags only in the fall test had the greatest impact on the numbers of students with very high positive growth. Likewise, Table 3 shows that removing students who triggered flags only in the spring test had the greatest impact on the numbers of students with very high negative growth. These results are consistent with our expectation that high positive and negative

growth scores are associated with low-ISV affecting only one of the two test events whose scores were used in measuring growth.

This effect is illustrated in Figure 1, which shows the scatterplots of the fall and spring reading RIT scores for the 9th graders in the sample both before and after the flagged students were removed. Although the upper display depicts a positive relationship, there were a number of outliers both above and below the major axis of the distribution. Outliers well above the major axis indicated substantial negative growth, with RIT growth scores as low as -71, while outliers well below the major axis indicated high positive growth, with scores as high as +84. As a reference, 9th graders taking the MAP in reading tend to show RIT fall-to-spring growth values in the single digits, with values rarely exceeding 20 points. Hence, the outliers in the upper display of Figure 1 indicate the presence of unreasonable growth scores.

The lower display depicts the relationship between the fall and spring RIT scores when the scores of students who triggered one or more effort flags were removed. It is evident that after removal the graph had few outliers, supporting the conclusion that the outliers were due to low student effort.

The impact of score removal can also be seen in the correlations between the 9th grade fall and spring RIT scores. The correlation increased from .82 to .86 after student removal, even though the standard deviations of the RIT scores decreased (from 16.06 to 14.84 in the fall and from 15.97 to 14.49 in the spring). Thus, student removal had the twin effects of reducing the range of scores for both RIT variables while increasing their correlation—a finding that is consistent with a conclusion that a source of CIV had been removed from the test data.

Table 4 shows the impact of low ISV on growth scores. For students triggering no flags, growth averaged 5.63 RIT points. For those students triggering effort flags in the fall only (whose growth scores would be positively biased), mean growth was roughly twice as high. For those triggering effort flags in the spring only (whose growth scores would be negatively biased), mean growth was markedly lower and slightly negative. In addition, the standard deviations of growth scores reflect the presence of effort-related CIV. The standard deviations were larger for students triggering flags during only one of their test events (relative to those triggering no flags), and larger still for those triggering flags in both fall and spring.

Finally, Table 5 shows the relationship between student growth and the number of effort flags triggered. For flags triggered only in the fall, there was a positive relationship, while for those triggered in the spring the relationship was negative. The spring findings indicate that the average RIT score for students exhibiting three flags in the spring was over 3 RIT points lower than it was in the fall, after a full year of instruction. This underscores the degree to which low effort can negatively bias proficiency estimates and distort assessment of growth.

Collectively, the evidence strongly indicated that the flagged test events reflect the presence of non-effortful test-taking behavior. Such behavior diminished the validity of the resulting RIT scores as estimates of student proficiency (which, in turn, invalidated growth scores based on the RIT scores).

Example 2: Predicting College Readiness

The second example, which further illustrates on the impact of low ISV on score validity, studied the relationships between MAP reading scores and several college readiness

assessments in reading (EXPLORE, PLAN, and ACT). The data from an earlier study (NWEA, 2011) were analyzed to better understand how low student effort affects the relationships between test scores. In that study, EXPLORE, PLAN, and ACT scores from 26,000 students were linked to their MAP assessment RIT scores from the same (or prior) testing season.

We focused only on the data from approximately 17,000 of the students whose MAP reading data could be linked to college readiness scores from the same testing season. Data from the MAP test events were subsequently evaluated using the three effort-related flagging criteria.

Because the MAP scores and the college readiness test scores are each measures of the same construct (reading proficiency), we would interpret sizable positive relationships among their scores as conventional evidence of test score validity. Table 6 shows the degrees of linear relationships between the MAP and the three college readiness tests, both before and after invalid MAP test scores were removed. To aid comparisons, the degree of linear relationship is expressed as a squared correlation (representing the proportion of shared variance between the two variables). For two of the three college readiness tests, the squared correlation increased after removal of invalid scores. The largest increase occurred with ACT reading, for which the highest percentage of students with invalid MAP scores was removed (17%).

Table 6 also shows that the squared correlations based only on the students with invalid MAP scores were much lower, consistent with the presence of construct-irrelevant variance in the MAP data. These squared correlations were not near zero, however, because students rarely behave non-effortfully throughout an entire test event. More typically, a

student will disengage during only a portion of a test event, suggesting that the scores from invalid test events reflect a mixture of effortful and non-effortful test taking behavior (Wise & DeMars, 2006). The weak, yet positive, correlations for invalid MAP scores observed in Table 6 would be expected under this conceptualization.

These results illustrate that the presence of scores with low ISV attenuates the degree of association between variables. In the present example, this decreased the ability of MAP scores to predict college readiness.

Summary

Whenever the purpose of measurement is to inform an inference about an individual student, it is important that we be able to trust that the student's test score is valid. Such confidence comes from the dual requirements that (a) we are using a competently-designed test can provide valid scores from the target population of students and (b) that the student actually responds to that test so as to permit a valid measure of his or her level of proficiency. While the first requirement pertains to traditional score validity, the second pertains to ISV. The central assertions of this paper are that (a) traditional validity evidence is necessary, but not sufficient, to ensure that an individual student's score is valid, and (b) an additional process is needed to gather ISV-related validation evidence.

Because ISV is the degree to which an individual's test score is free from sources of student-related CIV, the information to be collected focuses on indications that CIV has affected the test score. We described a six-step ISV process, in which the sources of CIV that represent the most serious threats to ISV are identified, indicators are established along with appropriate criteria, invalid test events are identified, policies are developed for

effectively dealing with invalid test events, and evidence is gathered that measurement is improved by identifying test events that reflect low ISV.

The most common sources of CIV that threaten ISV are familiar to test givers. They are student variables that include: low test-taking effort, high test-taking anxiety, fatigue, illness, and cheating. To identify these threats to score validity, there is an array of choices of indicators we might use. There may be occasions when student self-report scales could provide useful information (e.g., with effort or anxiety). In addition, there are a number of statistical indicators, such as person fit indices, response accuracy, or item response time (if the test is computer administered) that can be considered. Finally, there are different types of biometric information that can be used or might be used in the near future, such as galvanic skin response (GSR), heart rate, blood pressure, eye movements, or brain patterns. Of course, research will be needed to understand the most unobtrusive and nonreactive ways to collect such information and the probative value of such measures as indicators of ISV threats will need to be established.

Ultimately, we hope that consideration of ISV will become a routine aspect of educational measurement. Test givers can exert a great deal of control about the content of a test and the conditions under which the test is administered. It is important to recognize, however, the limits of the test giver's control. In practice, it is useful to be mindful of potential construct-irrelevant threats to score validity and the extent to which they can meaningfully distort test scores. Furthermore, it is important that we have methods for identifying the presence of such threats and policies in place for dealing with them. Like the school psychologist administering an individually administered test, we should be able to evaluate each test event for evidence that it has been affected by CIV, and that the resulting

test score cannot not be trusted to provide valid information about the student. It is through these considerations that we can ensure the validity of individual student scores.

References

Anastasi, A. (1988). *Psychological testing (6th Ed.)* New York: Macmillan.

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

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

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23(1),* 17-27.

Hauser, C., Kingsbury, G, G., & Wise, S. L. (2008, March). *Individual validity: Adding a missing link.* Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th Ed., p. 17-64). Westport, CT: Praeger.

Kingsbury, G. G., & Hauser, C. (2007, April). *Individual validity in the context of an adaptive test.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Messick, S. (1984). The psychology of educational measurement. *Journal of Educational Measurement, 21,* 215-237.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd, ed., p. 13-103). Washington, DC: American Council on Education and Macmillan.

Mislevy, R. J. (2009). Validity from the perspective of model-based reasoning. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications.* Charlotte, NC: Information Age.

Northwest Evaluation Association (2011, December). *A study of the alignment of the RIT scales of NWEA's MAP assessments with the college readiness benchmarks of EXPLORE, PLAN, and ACT.* Portland, OR.

Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement, 34,* 213-232.

Schnipke, D.L. & Scrams, D.J. (2002). Exploring issues of examinee behavior: Insights gained from response-time analyses. In C. N. Mills, M. T. Potenza, J. J. Fremer, & W. C. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments.* Mahwah, NJ: Lawrence Erlbaum Associates.

Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement, 43,* 19-38.

Wise, S. L., & DeMars, C. E. (2010). Examinee non-effort and the validity of program assessment results. *Educational Assessment, 15,* 27-41.

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18,* 163-183.

Wise, S. L., & Ma, L. (2012, April). *Setting response time thresholds for a CAT item pool: The normative threshold method.* Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, Canada.

Table 1

Numbers of Effort Flags Triggered at Each Time Period

Test Session	Flag	N	%
Fall, 2010	A (Overall RTE)	18,948	6.59
	B (Local RTE)	20,128	7.00
	C (Joint RTE and Accuracy)	12,452	4.33
	At Least One Flag	23,952	8.33
	Zero Flags	263,738	91.67
Spring, 2011	A (Overall RTE)	12,873	4.47
	B (Local RTE)	13,660	4.75
	C (Joint RTE and Accuracy)	8,110	2.82
	At Least One Flag	16,423	5.70
	Zero Flags	271,267	94.29

Note. N = 287,690.

Table 2

Decrease in the Numbers of Students from Various Growth Score Intervals When Low-ISV Test Events From the Fall Testing Were Removed

	Growth Score Interval[a]			
	0 to 10	10 to 20	20 to 30	> 30
All Students	128,898	61,936	12,051	2,479
If Students with Low ISV are Removed	5,710	5,095	2,281	1,016
Percent Removed	4.43%	8.23%	18.93%	40.98%

Note. Students triggering as least one effort flag during both the fall and spring testing sessions were excluded (n = 6,749). Students who exhibited low effort in the fall testing only would be expected to exhibit high positive growth.

[a] The standard error of growth scores in reading was approximately five points.

Table 3

Decrease in the Numbers of Students from Various Growth Score Intervals When Low-ISV Test Events From the Spring Testing Were Removed

	Growth Score Interval[a]			
	< -30	-30 to -20	-20 to -10	-10 to 0
All Students	232	929	8,839	65,577
If Students with Low ISV are Removed	171	411	1,375	2,992
Percent Removed	73.71%	44.24%	15.56%	4.56%

Note. Students triggering as least one effort flag during both the fall and spring testing sessions were excluded (n = 6,749). Students who exhibited low effort in the spring testing only would be expected to exhibit high negative growth.

[a] The standard error of growth scores in reading was approximately five points.

Table 4

Relationship Between Effort Flags and MAP Growth Scores

Flagged for Effort?	N	Mean Growth	SD Growth
No	254,064 (88%)	5.63	8.42
In Fall Only	17,203 (6%)	10.95	11.95
In Spring Only	9,674 (3%)	-0.27	12.52
In Both Fall and Spring	6,749 (2%)	3.99	14.84

Table 5

Mean Growth, by Number of Effort Flags Triggered, for Each Testing Period

Test Session	Number of Effort Flags	N	Mean	SD
Fall, 2010	0	263,738	5.41	8.68
	1	4,462	8.40	10.40
	2	6,522	10.02	11.75
	3	6,219	13.76	12.63
Spring, 2011	0	271,267	5.96	8.78
	1	2,757	2.64	10.95
	2	3,793	0.05	11.83
	3	3,124	-3.22	13.90

Note. 6,749 students triggering as least one effort flag during each of the testing sessions were excluded.

Table 6

Squared Correlations between MAP Reading Scores and College Readiness Test Scores

		Based on All Students		Based Only on Student Scores Classified as Valid		Based Only on Student Scores Classified as Invalid	
College Readiness Test	Grade	Squared Correlation	N	Squared Correlation	N	Squared Correlation	N
EXPLORE Reading	8	.43	11,055	.44	10,085	.14	970
PLAN Reading	10	.38	3,458	.38	3,122	.13	336
ACT Reading	11	.49	2,459	.53	2,046	.24	413

Note. A squared correlation indicates the proportion of variance shared between two variables.

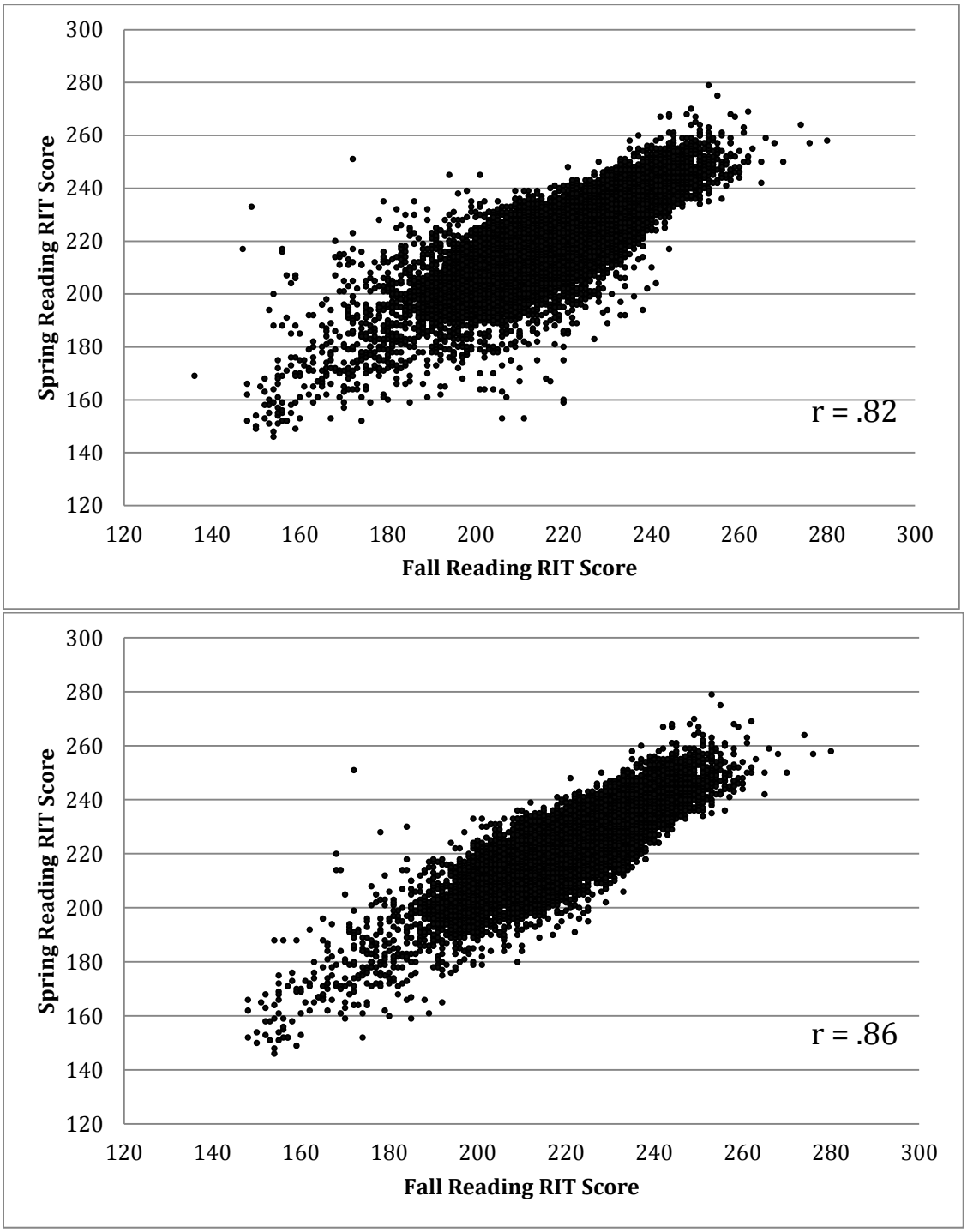


Figure 1. Scatterplot of the fall and spring reading RIT scores for 19,254 9th graders (top display) and after students triggering one or more effort flags on either test were removed (bottom display). A total of 4,289 students were removed (22%). There were 1,267students who exhibited low effort on both tests (7%).

Appendix

Effort Flagging Criteria for Identifying Low ISV Scores

When multiple-choice items are being administered on a computer-based test, there are two types of behaviors that indicate a student who has become disengaged from his test and is exhibiting non-effortful test taking behavior. First, he may respond to items very rapidly (i.e., faster than it should take him to read the item and thoughtfully consider the problem it poses). Second, his answers tend to be correct at a rate that is consistent with what would be expected by chance through random guessing. Rapid responses typically exhibit chance-level accuracy. Additionally, chance-level accuracy can provide supporting evidence of non-effortful test taking. Both response time and answer accuracy are used in the criteria for flagging MAP test events as exhibiting low ISV due to effort.

Rapid-guessing behavior is identified using *response time effort* (RTE; Wise & Kong, 2005), which is based on the conceptualization that each item response can be classified as reflecting either *rapid-guessing behavior* or *solution behavior* (Schnipke & Scrams, 1997, 2002). This classification is done using pre-established time thresholds for each item using the normative threshold method (Wise & Ma, 2012) set at 10 percent. This means that the threshold for an item is set at 10 percent of the average time students have historically taken to answer the item. RTE for a test event equals the proportion of the student's responses that were solution behaviors. This leads to the first effort flag:

Flag A: If the student gave rapid guesses to at least 15% of the items (overall RTE $\leq$ .85).

Flag A specifies the amount of rapid-guessing behavior that would be tolerated over the entire test event. Test-taking effort, however, is not all-or-none. Students sometimes

exhibit non-effort during only a portion of a test event. This complicates the identification of non-effortful behavior, and an overall indicator may not be sensitive to detecting lesser degrees of non-effort. For example, if a student gave good effort on the first 30 items of a 50-item CAT and then gave 7 rapid guesses during the remaining 20 items, his non-effort at the end of the test would not be enough to trigger Flag A.

One solution to this problem is to consider rolling subsets of the items from a test event. For example, for subsets of size 10, we would consider items 1-10, then 2-11, then 3-12, and so on, until the end of the test. In general, for a *k*-item test and subsets of size *r*, there will be $(k-r)+1$ rolling subsets. Using rolling subsets of size 10, we developed two additional RTE-based flags for considering low effort on a more local level:

Flag B: If the student exhibited low RTE (local RTE $\leq$ .70) on at least 20% of the rolling subsets.

Although item response time provides a useful indicator of non-effortful behavior, there are times when students become disengaged from during a test event without exhibiting rapid-guessing behavior. Inspection of MAP data indicates, for example, a number of test events in which a student gave low accuracy responses to items administered near those classified as rapid guessing. The combination of low accuracy responses and rapid guessing represents across a string of items represents a potent indicator of non-effortful behavior.

Low-accuracy responses should be evaluated carefully, however, because they could also be due to the student receiving items that were much too difficult for him. In principle, this alternative explanation should not pose a problem for a CAT because the CAT algorithm strives to select and administer items that a given examinee has a .50 probability of

answering correctly (which requires that the item pool is able to select items that are well targeted to each student). With MAP, however, it occasionally occurs that close targeting is not possible. This tends to happen in higher grades for high proficiency students, which means that items of high enough difficulty were not available to administer during those MAP test events[1].

Low accuracy responses should be used as indicators of low effort only if it is established that they were not due to poorly targeted items. To accomplish this, a pool adequacy requirement is imposed specifying that low response accuracy will only be considered for test events in which at least 60% of the time during the CAT, the student received an item whose difficulty was no more than three RIT points away from the student's momentary proficiency estimate[2]. This led to the development of a third effort flag based on the joint occurrence of rapid responses and low accuracy on *any* of the rolling subsets:

Flag C: If the student passed no more than two items (local accuracy ≤ .20) and gave three or more rapid guesses (local RTE ≤ .70) on any 10-item subset, and at least 60% of all of the administered items were within three RIT points of the student's momentary proficiency estimate.

In the effort analysis of MAP data, the score from a student's test event was classified as invalid due to low ISV if any one of the three effort flags was triggered.

---

[1] MAP is usually administered to students several times each academic year and, once an item has been administered to a student, that item cannot be re-administered to the same student for 24 months. After multiple MAP administrations to low proficiency students, this can lead to a shortage of available closely targeted items.

[2] The standard errors of student scores in math are typically about 3.0 RIT points, while those in reading are about 3.2 RIT points.

Identifying Non-Effortful Student Behavior on Adaptive Tests:

Implications for Test Fraud Detection

Steven L. Wise, Lingling Ma, Robert A. Theaker

Northwest Evaluation Association

Paper presented at the 2012 Statistical Detection of Potential Test Fraud Conference, Lawrence, Kansas

Abstract

Test fraud typically involves actions taken by individuals to positively distort a test score in such a way that it overstates what a student knows and can do. Teacher effectiveness, however, is increasingly being evaluated using student growth (i.e., the difference between scores at Time 1 versus Time 2). In this context, a novel type of potential fraudulent behavior becomes a concern. Test givers may be motivated to try to depress student test-taking effort at Time 1 in order to inflate the subsequent growth scores. To investigate the extent to which this represents a real practical problem, data were analyzed from a set of charter schools that has for years used fall-spring growth data to evaluate teacher effectiveness. The results showed clear evidence of lower effort in the fall, supporting the conclusion that this differential effort represents a real threat to the validity of inferences made on the basis of growth scores.

Identifying Non-Effortful Student Behavior on Adaptive Tests: Implications for Test Fraud Detection

The goal of educational assessment is to produce test scores that accurately indicate what students know and can do in a particular content domain. This goal is pursued through the administration of a test whose items adequately represent that content domain and whose items are sufficient in number to yield reliable scores. The scores from such a test are then used to either (a) make inferences about individual students (e.g., What is Michael's level of math proficiency?) or (b) make inferences based on the test scores of groups of students (e.g., How well did the students at Lakeside Middle School do on the state assessment in math?).

Obtaining a valid score for a particular student, however, requires more than just the administration of a well-developed standardized test. There are a number of potential threats to the validity of an individual test score. Some threats pertain to the behavior of the student (e.g., Did he give good effort to his test? Was he feeling ill? Did he cheat to try to get a good score?). Others pertain to the context in which the test occurred (e.g., Was the testing done late in the day? Were there noisy distractions?). Additionally, after the test event had concluded, errors might be made in scoring the test, which could result in a score that was either too high or too low.

Because these types of potential threats to validity are unrelated to the achievement construct under study, and because they affect some students more than others, they introduce construct-irrelevant variance into the test scores (Haladyna & Downing, 2004). *Individual score validity (ISV)* addresses the trustworthiness of individual test scores (Hauser & Kingsbury, 2009; Hauser, Kingsbury, & Wise 2008; Kingsbury &

Hauser, 2007). Wise, Kingsbury, and Hauser (2009) defined ISV as the extent to which test scores are free from construct-irrelevant factors. They encouraged measurement professionals to identify the most serious construct-irrelevant threats to ISV in a given measurement context, and to develop methods for assessing the degree to which those threats had affected particular scores.

An alternative definition for ISV, and one that is useful for this paper, is based on the concept of score *distortion*:

$$\text{Distortion}_i = \begin{bmatrix} \text{What the test score} \\ \text{indicates that student } i \\ \text{knows and can do} \end{bmatrix} - \begin{bmatrix} \text{What student } i \\ \text{actually knows} \\ \text{and can do} \end{bmatrix}. \tag{1}$$

Equation 1 is an abstract quantity, as we will not know for certain the student's actual level of proficiency. Conceptually, however, ISV for a student is attained when distortion of his score is equal to zero (or at least within the limits of measurement error). Positive values of distortion correspond to instances when the test score over-estimates the student's proficiency level; negative values indicate an under-estimation of proficiency. Positive and negative distortions correspond to different types of threats to ISV.

Positive distortion suggests the possibility of at least one of several types of test fraud initiated by the student, who will frequently be motivated to attain the highest score that he can. First, the student could have copied answers from a more proficient test taker during the test session. Second, he could have acquired pre-knowledge of at least some of the test items he would receive, and made sure he knew the answers to them prior to the test session. Third, he may have brought in notes or surreptitiously used technology (such as a smartphone) to acquire the answers to questions during the test session.

Because of the pressures of teacher and school accountability, test givers (such as teachers or principals) might also be motivated to engage in fraudulent testing practices. If specific items to appear on a test were known to test givers beforehand, they could provide that information directly to students before the testing session. During a testing session, they could subtly (or not) point out to students their incorrect answers and possibly indicate the answers they should be giving. After the testing session, test givers could privately alter students' answer sheets, either by filling in the correct answers to omitted items or by changing incorrect answers to correct.

Regardless whether it is initiated by the student or the test giver, test fraud is intended to create positive score distortion by producing test scores that overstate what students actually know and can do. Figure 1 illustrates these dual influences on distortion. The solid arrows indicate that students and test givers can directly induce distortion (and diminish ISV) through their actions.

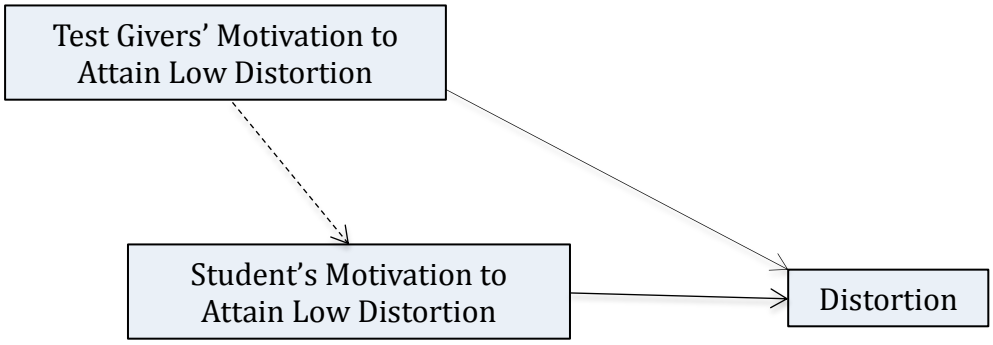


Figure 1. The intentional influences of students and test givers on distortion.

Negative distortion, in contrast, is attributable to the influence of construct-irrelevant factors leading to scores that underestimate what students know and can do.

Any factor that degrades test performance during the test session induces negative distortion. This can include, but is not limited to, student illness or emotional distraction, testing rooms that are too warm/cold or have noise distractions, or unmotivated students who do not give good effort throughout the test. Unlike positive distortion, negative distortion is generally not associated with intentional test fraud.

Test Fraud Through Negative Distortion

There is, however, at least one scenario in which test fraud *could* be induced through negative distortion. There recently has been a growing emphasis in U.S. schools on evaluating teacher effectiveness based in part on student growth data. Growth in this context is defined as the difference between a student's proficiency level at Time 1 and that at Time 2 (assuming that the two scores share a common measurement scale). The higher the levels of growth that a teacher's students exhibit, the more positive will be the teacher's evaluation. If a teacher can somehow depress his students' scores at Time 1, then student growth will consequently be inflated. That is, if the teacher can induce negative distortion at Time 1, then he will induce positive distortion on student growth scores.

How could this be accomplished? One way would be through the manipulation of student motivation. There is a well-documented relationship between test-taking motivation and performance; if a student is unmotivated and behaves in a non-effortful manner, the resulting test score is likely to underestimate his actual level of proficiency (Wise & DeMars, 2005). In testing situations where students perceive few or any personal consequences associated with test performance, they may be particularly responsive from cues from their teacher regarding how much effort they should expend. Thus, to the degree to which a teacher downplays the importance of the Time 1 test and emphasizes the

importance of the Time 2 test, positive distortion could be induced. The dotted line in Figure 1 indicates this indirect influence, by which the test giver induces distortion through manipulation of the student's motivation.

Students exhibiting markedly lower test-taking effort at Time 1 than at Time 2 would characterize the scenario described above. Thus, differential levels of effort provide an important indicator that distortion of growth scores has occurred.

Measuring Test-Taking Effort

As computer-based tests (CBTs) have become more common, there has been an increased interest in the uses of item response time (which can be collected during a CBT) to improve the measurement of academic achievement. One research theme has focused on the use of response time to investigate examinee engagement. Early research (Bhola, 1994; Schnipke & Scrams 1997, 2002) investigated, using item response time, changes in examinee behavior as time was running out during a speeded, high-stakes test. They found that many examinees switch strategies from trying to work out the answers to items (termed *solution behavior*) to rapidly entering answers to remaining items in hopes of guessing some of them correct (termed *rapid-guessing behavior*).

Wise and Kong (2005) observed that rapid-guessing behavior also occurred during unspeeded low-stakes CBTs. They showed that in this context rapid-guessing behavior indicates instances when a student was not expending effort toward attaining a good score[1]. Wise and Kong developed a measure of test-taking effort, termed *response time effort (RTE)*, which equals the proportion of a student's responses that were solution behaviors. An RTE

[1] Regardless of the stakes of the test or whether or not the test is speeded, a rapid guess indicates essentially the same thing—that the examinee was not engaged in solution behavior.

value of 1.0 indicates that a student exhibited only solution behavior, a value of .90 indicates 10% of the item responses were rapid guesses, and so on.

Identification of rapid-guessing behavior requires that a time threshold be established for each item. This permits each item response to be classified as either a solution behavior or a rapid guess. Two basic principles are followed in establishing time thresholds. First, we want to identify as many instances of non-effortful item responses as possible. Second, we want to avoid classifying effortful responses as non-effortful. There is a tension between the two principles such that the first encourages us to choose a longer threshold, while the second encourages us to choose a shorter one. Thresholds are chosen to balance the two principles, with the second principle being of higher priority. Good discussions of item threshold identification methods and issues are found in Ma, Wise, Thum, and Kingsbury (2011) and Wise and Ma (2012).

The identification of rapid-guessing behavior is important because it indicates the presence of item responses that exert a negative bias on a proficiency estimate. This is due to rapid guesses being correct at a rate that is usually markedly lower than what would have been the case had the student exhibited solution behavior. Therefore, the more rapid guesses occur during a test event, the more negative distortion is likely present in a test score. Thus, the presence of rapid-guessing behavior provides useful evidence that a score that has low ISV.

When a computerized adaptive test (CAT) is used, an additional indicator of low student effort is provided by the accuracy of the item responses. However, unlike rapid guessing—which can be assessed for each item response—accuracy must be considered across a set of items. The CAT algorithm is designed to select and administer items that a

student under solution behavior has about a .50 chance of getting correct. Under rapid-guessing behavior, in contrast, items will be correct at a rate consistent with random guessing. For multiple-choice items with five response options, a student would be expected to provide a correct response about 20% of the time. Hence, because responses to items administered in a CAT will have consistent, differential accuracy rates under solution behavior and rapid-guessing behavior, the accuracy of a student's responses to a set of items can be evaluated as to whether it appears to reflect solution behavior or rapid-guessing behavior. For example, if during the last half of a test a student passed only 22% of his items on a CAT, we might decide that he was not giving effort during that portion of the test and conclude that his score should be considered as reflecting low ISV.

Thus, both response time and response accuracy can provide valuable information about the effort that was expended by students during a CAT. In our research with NWEA's *Measures of Academic Progress (MAP)*, which is used to measure the academic growth of primary and secondary school students, we have developed a set of five flagging criteria for identifying test events that yield scores with low ISV. These heuristic criteria, which are described in the Appendix, are based on a combination of RTE and response accuracy, either singly or in combination. The criteria have been shown to identify many instances of non-effortful student behavior (Wise & Ma, 2012), and can be used to evaluate the degree of test-taking effort exhibited at different time periods.

A Case Study of Teacher Evaluation and Test-Taking Effort

It is one thing to describe a scenario in which test givers *could* manipulate student effort to inflate growth scores, and another to demonstrate that it represents a real problem. That is, if a problem has little chance of actually occurring, its solution may have

little practical value. To evaluate this issue, we studied data from a context in which student growth has been used as part of a teacher evaluation process for a period of time. These data should provide a basis to assess the degree to which Time 1 versus Time 2 effort discrepancies are present.

Data Source

This case study focused on a large charter school management organization that operates charter schools in multiple U.S. states. They have used MAP test results as part of their teacher evaluation system for a number of years. The evaluation system for teachers includes four components, with a substantial portion being an evaluation of student achievement based on fall-to-spring growth results for MAP.

The implementation of MAP as part of the teacher evaluation system is longstanding. As the charter school organization has increased its capacity around data use, it has refined and implemented more sophisticated approaches to measuring teacher effectiveness from test data. From the teachers' standpoint, however, one aspect has remained constant: a sizable portion of their evaluation is based on the amount of student academic growth in MAP that is observed between the beginning (i.e., fall) of an academic year and the end (i.e., spring) of that academic year. Hence, the evaluation system is consistent with the scenario described earlier in which there could be an incentive for test givers to try to depress fall scores with the goal of inflating growth scores. Occurrences of this could be identified by markedly more non-effortful test-taking behavior being observed in the fall as compared to the spring.

Student effort was measured using both RTE and the percentage of test events whose scores were classified as invalid using the five effort criteria described in the

Appendix. Ideally, we sought to study effort for groups of students defined by common teachers or classrooms. For our initial analyses, however, we were able to obtain reliable student groupings only at the school level. Nevertheless, we believed that we might observe differential fall-spring effort even at that coarser level.

The data analyses focused on MAP fall-spring growth scores in math and reading in grades 3-8 from 39 charter schools within a single U.S. state. MAP tests are untimed, interim CATs, with the tests in math being generally 50 items in length, while those in reading are generally 40 items in length. MAP proficiency estimates are expressed as scale (RIT) scores on a common scale that allows growth to be assessed as students are tested at different times. The standard errors of the fall-spring scores in math are typically about 4.25 points, while those in reading are about 4.50 points.

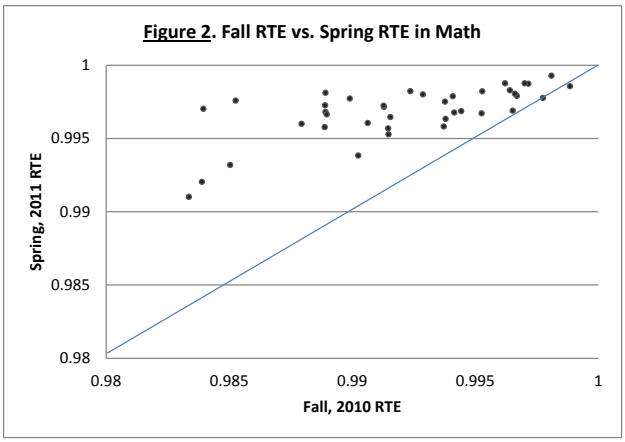
Results and Discussion

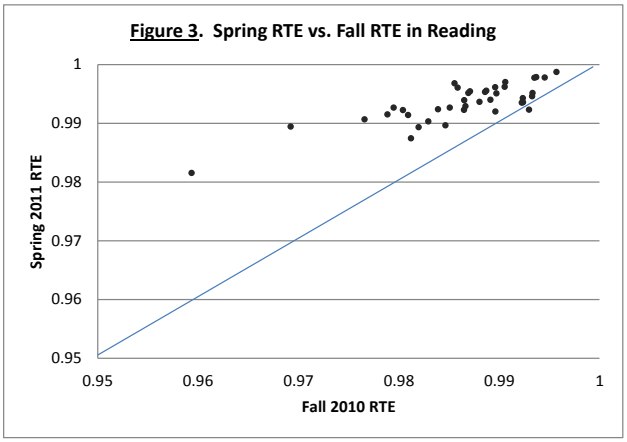
Table 1 shows descriptive statistics for all of the students across the charter schools. Two findings are particularly noteworthy. First, both mean RTE and percent invalid scores indicate that non-effortful behavior occurred more frequently in the fall than in the spring. Second, non-effortful behavior was markedly more prevalent in reading than in math. The first finding was reported by Wise, Ma, Kingsbury, & Hauser (2010) and the second has been reported in two previous studies (Wise et al., 2009; Wise et al., 2010).

Table 1. Descriptive Statistics for RTE, Percentage of Invalid Scores, and Fall-to-Spring Growth

		Mean RTE		Percentage of Invalid Scores			
Content Area	N	Fall	Spring	Fall	Spring	RIT Growth Mean	RIT Growth SD
Math	13,416	.993	.997	3.4	1.4	10.39	7.75
Reading	13,463	.987	.994	7.3	4.1	8.17	8.81

The accuracy of solution behaviors and rapid-guessing behaviors was consistent with our expectations with data from an adaptive test. The accuracy rates for item responses classified as solution behaviors in math and reading were 51.0% and 51.3%, respectively. The accuracy rate for rapid guesses in math (whose items had 5 response options) was 21.0%, which is close to the value expected by random responding. Similarly, the accuracy rate for rapid guesses in reading (whose items had 4 response options) was 26.3%.

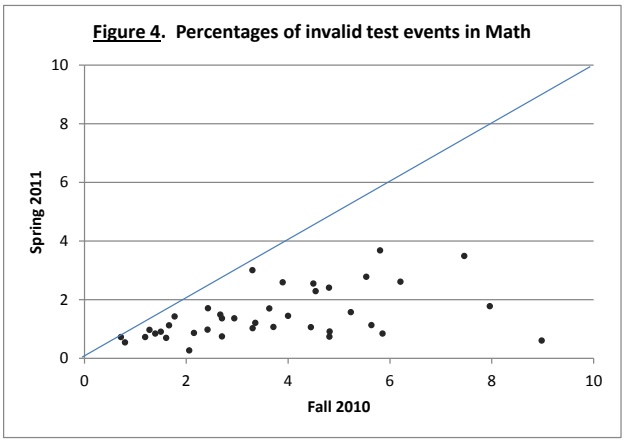


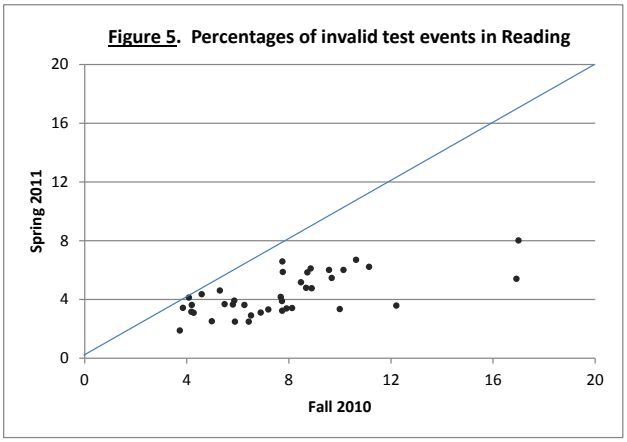


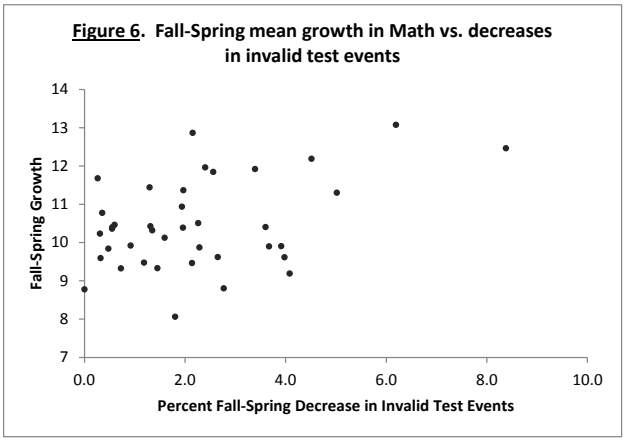
A scatterplot of spring versus fall mean RTE values in math is shown in Figure 2. If effort had been similar in the fall and spring at each school, the data points would be expected to fall near the main diagonal of the graph. Figure 2 shows, however, a somewhat surprising data trend indicating lower mean RTE in the fall for nearly every school. The results for reading, shown in Figure 3, indicate the same basic trend. In addition, there was substantial variation in the fall RTE values, with the graphs showing some schools with markedly lower fall RTE.

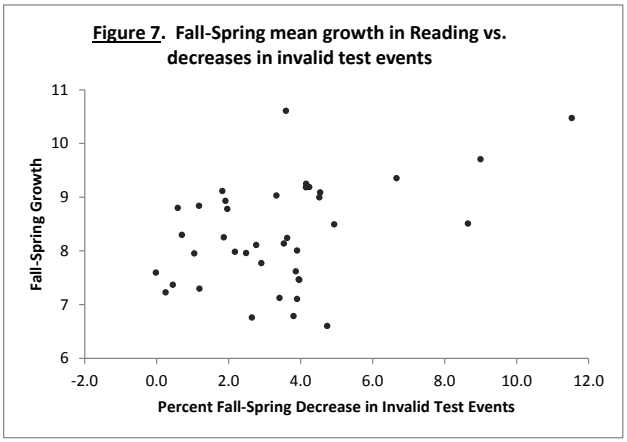
The percentages of invalid test events in math and reading are shown, in Figures 4 and 5, respectively. Again, similar results were found for each content area. There were

clear indications that effort was lower in the fall for virtually every school. In math, the difference in the preponderance of invalid test events was as high as 8 percentage points. In reading, the difference for one school exceeded 11 percentage points. These results suggest that lower fall effort was the norm rather than the exception in both math and reading.









The relationship between the fall-spring decrease in invalid test events and fall-spring growth is shown in Figure 6 for math and in Figure 7 for reading. In both content areas, there was an association between the two variables. Each of the three schools exhibiting a decrease in invalid test events that exceeded 5 percentage points showed

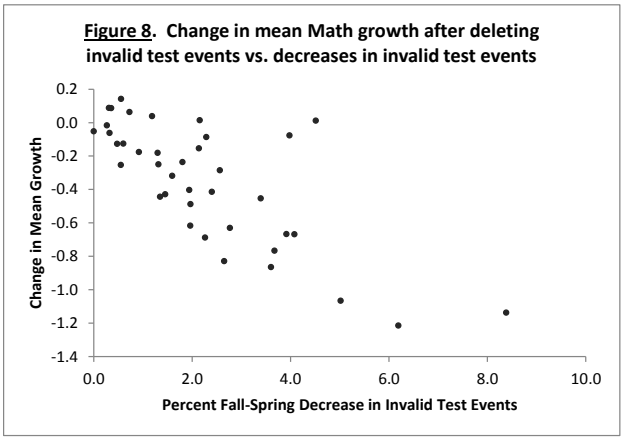
above average growth in math. Similarly, above average growth in reading was exhibited by all four of the schools showing at least a 5-point decrease in invalid test events.

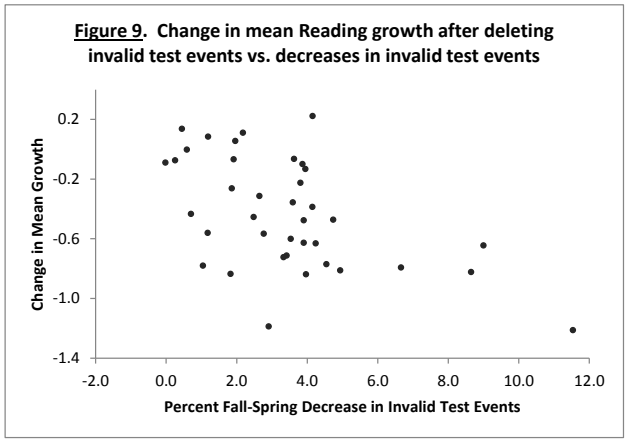
The results shown in Figures 4-7 provide the clearest answer to the general research question that this case study was intended to answer (i.e., is there evidence of disparate fall-to-spring student effort that induced positive distortion on growth scores?). The finding that nearly all of the schools showed lower effort in the fall suggests that (a) there may be a general tendency of test givers to downplay the relative importance of fall testing or (b) students may be relatively unmotivated to devote strong test-taking effort early in an academic year. Whichever explanation is more valid, the finding that the largest decreases in effort was associated with above average growth suggests that the degrees of differential fall-spring effort observed in these schools may have a meaningful impact on mean growth scores.

Demonstrating that differential effort can lead to positive distortion leads to a question regarding the degree to which this distortion biased the growth score estimates. Figures 8 and 9 show the impact on mean growth of deleting the scores from test events identified as invalid. There is a negative relationship, particularly for math. High degrees of differential effort appear to be associated with larger decreases in mean growth. These findings support the claim that higher growth is due, at least in part, to higher differential effort.

It might be tempting to interpret the vertical scales in Figures 8 and 9 (change in mean growth) as indicators of the amount of distortion induced by differential effort for a given school. There are, however, two caveats to consider. First, Figures 8 & 9 show what would happen if the scores from invalid test events were deleted. We do not know for

certain the degree to which the deleted scores were distorted by differential effort. Second, we are not asserting that only the scores from deleted test events were subject to distortion. Our effort flagging criteria are conservative in detecting non-effort, which means that we will classify a test event as invalid only if there is clear evidence of non-effortful behavior (Wise & Ma, 2012). This suggests the possibility of test events with lesser amounts of non-effortful behavior that would be undetected by our criteria.





The data analyses from the case study illustrate that differential effort does appear to occur in practice at the school level, and that this induces meaningful amounts of positive distortion in growth scores. This implies that test fraud could successfully be accomplished by test givers through a manipulation of the amount of relative effort students give to their fall and spring assessments.

General Discussion

This paper introduced and discussed the concepts of test score distortion. It was shown that most types of cheating (by either test givers or students) are intended to induce positive distortion on test scores, with the goal of overstating what students actually know and can do. Negative distortion, in contrast, is generally caused by the influences of one or

more sources of construct-irrelevant variance that exert a negative bias on test scores. When student growth is considered, however, it is important to realize that positive distortion on growth scores could be induced through a negative distortion of Time 1 scores through a manipulation of student effort.

How, specifically, could this be done? Given the high-stakes nature of many assessments, it might appear difficult to lower a student's test-taking effort. It should be noted, however, that many assessments deemed "high-stakes" are actually "little-to-no-stakes" from the perspective of the student. For example, the scores from state assessments given according to the No Child Left Behind legislation have carried enormous consequences for school accountability. From the student's perspective, in comparison, the scores from these assessments have carried few personal consequences. The student's grades are unaffected by performance on the state assessment, and scores on the assessment usually are unavailable until months after the test event.

Given the absence of personal (i.e., external) consequences associated with many standardized assessments, one might reasonably turn the question around and ask why students give *any* effort to these assessments. In the absence of personal consequences, a student's test-taking motivation is driven by internal factors, such as a desire to please teachers and parents, academic citizenship, competitiveness, and ego satisfaction (Wise & Smith, 2011). In this context, cues from the teacher about the importance of a particular test and the amount of effort expected from students can have a strong influence on the amount of subsequent effort that occurs. Certainly, students are generally not eager to take tests, and it may require only subtle cues from the teacher that a test event is not highly important for them to decrease the amount of effort expended. Teachers are used to

devoting a lot of time encouraging their students to give their best efforts to standardized assessments, and it require little more than inaction on their part during fall testing to induce a sizable positive distortion on fall-spring growth scores.

Nevertheless, it is important to consider what constitutes test fraud. In this paper, test fraud means that someone acted intentionally to induce a positive distortion on test scores. A teacher, however, who de-emphasizes the importance of fall assessment performance relative to the importance emphasized for spring testing may not view his actions as fraudulent. A principal who encourages her teachers to view the fall testing merely as a baseline (consequently de-emphasizing its importance), and thereby indirectly diminishing test-taking effort might not view her actions as fraudulent. But from the standpoint of score distortion, these actions constitute fraudulent behavior even if test givers do not view it as such.

Given the emerging emphasis of student growth data on teacher evaluations, it is likely that the problem discussed in this paper will emerge as a serious threat to the validity of inferences made about teacher performance. Teachers will increasingly be tempted by the evaluative benefits of discouraging effort on fall assessments. Moreover, teachers who would normally resist the temptation may adopt the practice anyway, because they believe that other teachers are doing so. That is, they may be trying to re-level what they perceive to be an uneven playing field. The pervasive differential effort exhibited by schools in the case study may be evidence that this has occurred in the charter schools. More research, however, is needed to understand the degree to which lower effort in fall testing is a naturally occurring phenomenon. A better understanding of normative

student effort behavior will allow us to more clearly identify occasions when test givers have intentionally manipulated their students' levels of effort.

Recommended Solutions

Student growth data provides valuable information to teachers about the instructional needs of individual students. That is its primary purpose. When growth scores are also used to evaluate teacher effectiveness, there is a very real risk that test givers will attempt to distort the growth scores so as to put themselves in the best possible evaluative light. In the process, however, they undermine the validity of the primary instructional use of the growth data. It is therefore important that this threat to growth score validity be effectively addressed. This can be done through a combination of detection, deterrence, and mitigation.

When a CAT is used, the five effort flagging criteria used in this paper can be useful in detecting scores that are invalid due to low ISV. This would permit one to audit the results from a school or classroom. If markedly lower effort was exhibited in the fall testing term, an investigation could be initiated into the causes of the discrepant effort, and whether or not those causes constitute test fraud. If a non-adaptive CBT is used, then the three flagging criteria that are not based on score accuracy alone could be used to detect invalid scores—though probably not as effectively as when a CAT is used. When paper-and-pencil tests are used, it would be more difficult to detect low effort. Post-test self-report measures regarding effort might be used, but the validity of their scores might be questionable because it would be difficult to ascertain how truthfully students would respond. This, then, represents an additional advantage of computer-based tests over paper-and-pencil tests.

If procedures were put into place to detect differential effort, then test givers should be made aware of them, as well as of potential sanctions that could result from their occurrence. The knowledge that differential effort could be detected should serve as a strong deterrent. This awareness should be augmented with professional development for test givers that clearly defines this type of test fraud and what constitutes acceptable and unacceptable practices.

The remaining problem concerns how to effectively deal with instances of differential effort that do occur. Simply deleting the growth scores based on invalid test events is unsatisfactory for two reasons. First, we would like to get valid instructional information for as many students as possible, and deleted scores would not provide this information. A second, more cynical reason is that the deletion of scores from invalid test events could lead to different fraudulent strategy. A teacher may be motivated to selectively discourage effort only from the students who he believes are likely to show the lowest growth. Strategic deletion of their data would provide an additional way for a teacher to inflate his students' growth.

To mitigate the problem of low student effort, statistical methods may be useful in both salvaging growth scores for as many students as possible and in facilitating more valid assessment of teacher effectiveness. Wise and DeMars (2006) investigated a, IRT-based method for adjusting scores for the effects of non-effortful test-taking behavior. This method involves classifying each item response as solution behavior or a rapid guess, and basing proficiency estimation for a student only on solution behaviors (i.e., ignoring rapid guesses). This method is based on the assumption that for a test event containing both solution behaviors and rapid guesses, the solution behaviors all reflect effortful behavior.

Bowe, Wise, and Kingsbury (2011) found evidence that this assumption does not always hold. Thus, additional research is needed to develop methods for statistically managing the effects of low student effort on test scores.

References

Bhola, D. S. (1994). *An investigation to determine whether an algorithm based on response latencies and number of words can be used in a prescribed manner to reduce measurement error.* Unpublished doctoral dissertation, University of Nebraska-Lincoln.

Bowe, B., Wise, S. L., & Kingsbury, G. G. (2011, April). *The utility of the effort-moderated IRT model in reducing negative score bias associated with unmotivated examinees*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23(1),* 17-27.

Hauser, C., Kingsbury, G, G., & Wise, S. L. (2008, March). *Individual validity: Adding a missing link.* Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

Hauser, C., & Kingsbury, G, G. (2009, April). *Individual score validity in a modest-stakes adaptive educational testing setting.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Kingsbury, G. G., & Hauser, C. (2007, April). *Individual validity in the context of an adaptive test.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Ma, L., Wise, S. L., Thum, Y. M., & Kingsbury, G. G. (2011, April). *Detecting response time threshold under the computer adaptive testing environment.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

Schnipke, D. L. & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Education Measurement*, *34(3),* 213-232.

Schnipke, D.L. & Scrams, D.J. (2002). Exploring issues of examinee behavior: Insights gained from response-time analyses. In C. N. Mills, M. T. Potenza, J. J. Fremer, & W. C. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments.* Mahwah, NJ: Lawrence Erlbaum Associates.

Wise, S. L., & DeMars, C. E. (2005). Examinee motivation in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10,* 1-18.

Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement, 43,* 19-38.

Wise, S. L., Kingsbury, G. G., & Hauser, C. (2009, April). *How do I know that this score is valid? The case for assessing individual score validity.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego.

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*, 163-183.

Wise, S. L., & Ma, L. (2012, April). *Setting response time thresholds for a CAT item pool: The normative threshold method.* Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, Canada.

Wise, S. L., Ma, L., Kingsbury, G. G., & Hauser, C. (2010, May). *An investigation of the relationship between time of testing and test-taking effort.* Paper presented at the annual meeting of the National Council on Measurement in Education, Denver, CO.

Wise, S. L., & Smith, L. F. (2011). A model of examinee test-taking effort. In J. A. Bovaird, K. F. Geisinger, and C. W. Buckendal (Eds.) *High-stakes testing in education: Science and practice in K-12 settings (pp. 139-153).* Washington, DC: American Psychological Association.

Appendix

Effort Flagging Criteria for Identifying Low ISV Scores

When multiple-choice items are being administered, there are two types of behaviors that indicate a student who has become disengaged from his test and is exhibiting non-effortful test taking behavior. First, he may respond to items very rapidly (i.e., faster than it should take him to read the item and thoughtfully consider the problem it poses). Second, his answers may be correct at a rate that is consistent with what would be expected by chance through random guessing. Rapid responses typically exhibit chance-level accuracy. Additionally, however, chance-level accuracy can sometimes occur in the absence of rapid responding. For these reasons, both response time and answer accuracy are used in the criteria for flagging MAP test events as exhibiting low ISV due to effort.

Rapid-guessing behavior is identified using *response time effort* (RTE; Wise & Kong, 2005), which is based on the conceptualization that each item response can be classified as reflecting either *rapid-guessing behavior* or *solution behavior* (Schnipke & Scrams, 1997, 2002). This classification is done using pre-established time thresholds for each item using the normative threshold method (Wise & Ma, 2012) set at 10 percent. This means that the threshold for an item is set at 10 percent of the average time students have historically taken to answer the item. RTE for a test event equals the proportion of the student's responses that were solution behaviors. This leads to the first effort flag:

Flag A: If the student gave rapid guesses to at least 15% of the items (overall RTE ≤ .85).

Flag A specifies the amount of rapid-guessing behavior that can be tolerated over the entire test event. Test-taking effort, however, is not all-or-none. Students sometimes exhibit non-effort during only a portion of a test event. This complicates the identification of non-effortful behavior, and the overall indicators may not be sensitive to detecting lesser degrees of non-effort. For example, if a student gave good effort on the first 43 items of a 50-item CAT and then gave rapid guesses to the remaining items, his non-effort on the last 7 items would not be enough to trigger Flag A.

One solution to this problem is to consider rolling subsets of the items from a test event. For example, for subsets of size 10, we would consider items 1-10, then 2-11, then 3-12, and so on, until the end of the test. In general, for a *k*-item test and subsets of size *r*, there will be $(k-r)+1$ rolling subsets. Using rolling subsets of size 10, we developed two additional RTE-based flags for considering low effort on a more local level:

Flag B: If the student exhibited low RTE (local RTE ≤ .70) on at least 20% of the rolling subsets.

Item response time is useful for identifying rapid–guessing behavior. Inspection of MAP data indicates, however, instances in which a student exhibited low accuracy in the absence of rapid guessing. This suggests that some students can become disengaged from during a test event without resorting to rapid-guessing behavior.

Low-accuracy responses should be evaluated carefully, because they could also be due to the student receiving items that were much too difficult for him. In principle, this alternative explanation should not pose a problem for a CAT because the CAT algorithm strives to select and administer items that a given examinee has a .50 probability of

answering correctly (which requires that the item pool is capable of providing items that are well targeted to each student). With MAP, however, it occasionally occurs that close targeting is not possible. This tends to happen in lower grades for very low proficiency students, which means that items of low enough difficulty were not available to administer during those MAP test events[2].

Low accuracy responses should be used as indicators of low effort only if it is established that they were not due to poorly targeted items. To accomplish this, a pool adequacy requirement is imposed specifying that low response accuracy will only be considered for test events in which at least 60% of the time during the CAT, the student received an item whose difficulty was no more than three RIT points away from the student's momentary proficiency estimate[3]. This led to the development of two additional flags related to response accuracy:

Flag C: If the student passed fewer than 30% of the items (overall accuracy ≤ .30) and at least 60% of all of the administered items were within three RIT points of the student's momentary proficiency estimate.

Flag D: If the student exhibited low accuracy (local accuracy ≤ .20) on at least 20% of the rolling subsets and at least 60% of all of the administered items were within three RIT points of the student's momentary proficiency estimate.

---

[2] MAP is usually administered to students several times each academic year and, once an item has been administered to a student, that item cannot be re-administered to the same student for 24 months. After multiple MAP administrations to low proficiency students, this can lead to a shortage of available closely targeted items.

[3] The standard errors of student scores in math are typically about 3.0 RIT points, while those in reading are about 3.2 RIT points.

Finally, the joint occurrence of rapid responses and low accuracy on any of the rolling subsets was considered to be particularly indicative of low effort. This led to the final effort flag:

Flag E: If the student passed no more than two items (local accuracy ≤ .20) and gave three or more rapid guesses (local RTE ≤ .70) on any 10-item subset, and at least 60% of all of the administered items were within three RIT points of the student's momentary proficiency estimate.

In the effort analysis of MAP data, a student's test event was classified as invalid on the basis of low ISV if any one of the five effort flags was triggered.

System and Bandwidth Requirements

QUICK REFERENCE

The following table lists the supported hardware and software components for using the Web-based MAP® system.

> **Important! Due to incompatibilities in hardware and software from third-party vendors, computers with PowerPC® processors can no longer be used for Web-based MAP testing after July 20, 2012.**

In most cases, the requirements are the same for both the MAP Administration and Reporting Center (MARC) and the MAP Student Testing Center (MSTC). Any differences are noted below.

Category	PC Requirements	Macintosh® Requirements
Operating systems and software		
Operating system	Microsoft® Windows XP® with Service Pack 3, Windows Vista®, or Windows® 7	Mac OS® X v10.5, v10.6, or v10.7
Browser for MARC*	Internet Explorer® 7, 8, or 9† or Firefox® 13	Safari® 4, 5, or 5.1 or Firefox 13
Browser for MSTC	MAP lockdown browser# for PC with Internet Explorer 7, 8, or 9	MAP lockdown browser# for Mac® with Safari 4, 5, or 5.1
Multi-media player	**Adobe® Flash® Player 10.2** or higher	
Reports viewer for MARC	PDF viewer, such as Adobe Reader®	
Computer hardware‡		
Screen resolution	1024 x 768 minimum	
Color depth	16-bit (32-bit recommended)	
Peripherals	Keyboard and mouse Headphones for MAP for Primary Grades (MPG) testing	

*MARC requires specific browser settings, such as allowing cookies and pop-ups for the MAP URL. For details, see the *System Administration Guide*.

†Any Internet Explorer 9 issues with finding saved testing sessions may be resolved by enabling Compatibility View.

#The MAP system requires a MAP lockdown browser for all testing. The lockdown browser reduces distractions to students during testing by restricting access to other applications and Web sites.

‡Computer hardware components, such as the processor and RAM, must meet the requirements to run the software listed above. Specifically, verify that the hardware meets the operating system and Adobe Flash Player requirements listed on the Web sites of the respective software vendors.

Bandwidth Availability

The key bandwidth requirement is available Internet bandwidth sufficient for adequate system performance during student testing.

The most important bandwidth considerations are:

- Number of computers available for concurrent tests
- Number of students taking MAP for Primary Grades (MPG) tests during the scheduled test window
  Note: MPG tests require the most bandwidth.
- Bandwidth available through your main Internet connection
- Bandwidth available through any limitation points in your internal network (such as school-to-district connections when the district has the main Internet connection)

The following table can help you determine the approximate bandwidth to allocate based on the number of concurrent testers in your organization.

Students Testing Concurrently	Average Available Bandwidth Required*	
	MAP Tests	MPG Tests†
50	0.15 Mbps	0.9 Mbps
100	0.3 Mbps	1.8 Mbps
250	0.75 Mbps	4.5 Mbps
500	1.5 Mbps	9 Mbps
1000	3 Mbps	18 Mbps

*The table identifies the average bandwidth in Megabits per second (Mbps) required to provide adequate performance and avoid system time-outs during the question-and-answer portion of student tests.

†MPG tests require more bandwidth because they include audio support and additional interactive features for young students. Since the grades taking MPG versus MAP tests may vary from term to term, the bandwidth requirements may likewise vary.

Additional bandwidth is needed at the beginning of each test to load the initial test questions and the question display software into cache in each computer's browser. The initial load is approximately 2.2 MB, and the load time varies based on available bandwidth. As an example, with available bandwidth of 0.6 Mbps for a single student, the initial load would take about 30 seconds.

To perform an Internet speed test, use a Web site such as www.speedtest.net and select Portland, Oregon as the location. Perform the test multiple times from each location where testing will take place, such as each school building. Test the speed at different times during the school day to understand the variations in bandwidth availability based on staff and student usage. Record the download speeds and use them to approximate your available bandwidth during testing.

The bandwidth requirements listed in this section are subject to change. In addition, the bandwidth you need may vary from one term to the next, and from year to year, based on your testing plans.

© 2011—2012 Northwest Evaluation Association
Measures of Academic Progress, MAP, and DesCartes: A Continuum of Learning are registered trademarks of NWEA in the United States or other countries. The names of other companies and their products mentioned in this documentation are the trademarks of their respective owners.

NWEA
Northwest Evaluation Association
Partnering to help all kids learn®

Measures of Academic Progress® (MAP®) Administration Technical Requirements: Client-Server MAP®

The following technical requirements for MAP testing have been created as a guide to ensure that you have a successful testing experience with your students. Please review the following information as you assess your technology and plan for conducting these assessments.

Your technology will need to support these major components:

- The Network Test Environment (NTE) folder
- One or more NTE Administration Workstations
- TestTaker Client Workstation
- Northwest Evaluation Association™ (NWEA™) Reports Site
- Wireless Requirements (Optional)

Network Test Environment (NTE) Folder Server

The NTE Folder is a folder that stores and hosts student and test database information. Students will use the TestTaker application to interact with the NTE folder—TestTaker gets test questions from the NTE, and sends each unique student response back to the NTE for recording along with other measurement data. The NTE Folder is a folder that can exist on any shared network resource accessible to the local user accounts on the testing workstations. Please note: It is strongly recommended that the NTE folder be hosted on server-class hardware in the same physical building and on the same local network as the TestTaker client workstations. The quality of connectivity between the server and the workstations, as well as the specifications of the server hosting the NTE (including the CPU, RAM, and available disk space) all contribute to the overall test-system performance. The minimum system requirements for the NTE Server are:

Minimum NTE Server Requirements		
CPU	Pentium® II (266MHz)	1. These requirements assume the NTE server is dedicated to MAP while testing is in progress. 2. Server hardware should include 256MB RAM per lab or class of 30 TestTaker client workstations. 3. Some tests require audio that can
RAM	256 MB	
Available Disk Space	600 MB (1.5 GB for MPG Audio)	
Operating System*	Novell® 4.0; Mac® OS X® Server 10.4 or	

©2012 Northwest Evaluation Association™

Minimum NTE Server Requirements		
	higher; Windows® 2000 Server; Windows 2000 Advanced Server; Windows Server 2003 Windows Server 2008	be downloaded independently from the test. 4. Operating systems should be server-class operating systems. Using desktop systems in a server role can dramatically reduce performance and may limit the number of concurrent connections allowed[1]. 5. Earlier AppleShare versions do not support enough open shares for MAP system usability.
Network	Windows 2000 Server; Mac using SMB; Mac G3 (with AppleShare® 6.3) Novell 4.0; Mac using AFP	
Anti-Virus	Any anti-virus package must have an exclusion to prevent it from actively scanning any NTE folder during testing.	

** OS X version 10.7.3 or higher is required for NTE data hosted on OS X Lion (10.7) servers*

NTE Admin 2 Workstation

The Network Test Environment (NTE) Admin 2 software is a Windows-based application used to perform the download and upload of tests, test results and student data to and from the NTE folder. The NTE Admin 2 workstation must have Internet access, and the local Windows account used to run NTE Admin 2 must have *Read* and *Write* access to all files and subfolders of the NTE folder. Additionally, the NTE Admin 2 software requires the Microsoft® .NET Framework (version 4 or higher) to be installed. If your computer does not already have the .NET Framework, you will be prompted to install it automatically. The minimum system requirements for running NTE Admin 2 are:

NTE Admin 2 Workstation Requirements	
CPU	1GHz
RAM	512MB
Operating System	Windows XP SP3 Windows Vista® SP1 or later Windows 7 Windows Server® 2003 SP2 Windows Server 2008 (not supported on Server Core Role) Windows Server 2008 R2 (not supported on Server Core Role) Windows Server 2008 R2 SP1

[1] Non-server versions of Microsoft Windows limit user connections to 10-20.

©2012 Northwest Evaluation Association™

TestTaker Client Workstation

All NWEA assessments are presented to students using TestTaker. The TestTaker application can be installed locally on testing workstations, or on a Windows-based server. In the case of server-based installations, Windows-based clients can be given a shortcut that points to the TestTaker installation on the server. Macintosh clients must have the TestTaker application locally installed. The table below lists the minimum TestTaker client workstation system requirements for optimal performance. TestTaker may work on other configurations, however, NWEA has not tested all possible configurations and recommends following these guidelines whenever possible. If you are planning on using an untested configuration, NWEA strongly recommends you spend some time evaluating if the software will operate reliably on that configuration.

Minimum TestTaker Client Workstation Requirements	
CPU	Pentium II (266MHz); Mac G4
RAM	256MB
Operating System	Windows 2000; Windows XP; Windows Vista; Windows 7; Mac OS X 10.5 – 10.7
Network	Windows 2000 Server; Novell 4.0; Mac using AFP 2.x-3.x
Available Disk Space	70 MB
Screen Resolution	Preferred: 1024 x 768 Minimum: 800 x 600
Mac Fonts	Mac systems must have Courier and Symbol fonts installed. For the correct display of MAP for Primary Grades tests, font smoothing must be enabled for all Mac systems used as TestTaker client workstations (see www.nwea.org/mpginfo for more information).
Adobe® Flash®	A custom version of Adobe Flash is embedded within TestTaker. Most MAP assessments use this embedded version; therefore, Adobe Flash must not be blocked on the network or client machines.
Sound Card	Required (for MAP for Primary Grades only)
Headphones	Required (for MAP for Primary Grades only)
Permissions	Each client accessing the installation must have full read/write/create permissions set to the NTE.
TestTaker Version	MAP assessments require the use of TestTaker version 8.0 or later.

©2012 Northwest Evaluation Association™

Wireless Recommendations (Optional)

The following are the minimum recommendations for testing in a wireless environment (LAN only) and assume no other activity on a wireless access point. NWEA requires a persistent connection to the wireless access point, free of interruptions, to successfully run TestTaker. Any outages in the connections, regardless of how brief, may cause errors during testing or require re-testing particular students.

Wireless Requirements	
Wireless Standard	802.11 G (54 mb/s) or higher
# concurrent connections	12-15 connections per wireless access point (20 max)

NWEA™ Reports Site

The NWEA Reports site is a web portal providing access to NWEA assessment tools that includes *Descartes: A Continuum of Learning*®, *Primary Grades Instructional Data*, *Dynamic Reporting Suite (DRS)*, classroom resources, growth reports, and operational reports. The Reports site is not browser specific to view online reports, however for optimal viewing NWEA recommends the following guidelines:

NWEA™ Reports Site Requirements	
Browser	In order to upload files, you must use Internet Explorer® as the upload functionality leverages IE's secure transfer of files. However, the following browsers support browsing the NWEA Reports Site: ▪ Internet Explorer ▪ Mozilla Firefox® ▪ Safari®
Crystal Reports	End of Term reports that contain assessment results from the district and school-level can be accessed using the Crystal Reports Viewer. This viewer is designed to run on a Windows OS®, and can be downloaded from the NWEA Reports Site.

©2012 Northwest Evaluation Association™