Creating a K-12 Adaptive Test: Examining the Stability of Item Parameter Estimates and Measurement Scales

G. Gage Kingsbury

Steven L. Wise

Northwest Evaluation Association

Reprint requests to:

G. Gage Kingsbury

Northwest Evaluation Association

121 NW Everett

Portland, OR 97209

Running Head:

Creating a K-12 adaptive test

Conflict of Interest Disclosure: This research was solely funded by the Northwest Evaluation Association, which employs both authors.

Abstract

Development of adaptive tests used in K-12 settings requires the creation of stable measurement scales to measure the growth of individual students from one grade to the next, and to measure change in groups from one year to the next. Accountability systems like No Child Left Behind require stable measurement scales so that accountability has meaning across time. This study examined the stability of the measurement scales used with the Measures of Academic Progress. Difficulty estimates for test questions from the reading and mathematics scales were examined over a period ranging from 7 to 22 years. Results showed high correlations between item difficulty estimates from the time at which they where originally calibrated and the current calibration. The average drift in item difficulty estimates was less than .01 standard deviations. The average impact of change in item difficulty estimates was less than the smallest reported difference on the score scale for two actual tests. The findings of the study indicate that an IRT scale can be stable enough to allow consistent measurement of student achievement.

Creating a K-12 Adaptive Test: Examining the Stability of Item Parameter Estimates and Measurement Scales

Developing an adaptive test to measure the achievement of K-12 students is an interesting challenge. Among the issues to be addressed are the ages of the students, the environment in the schools, the technology differences among schools, differences in curriculum and instruction, and differences in student capabilities as learning occurs. Approaches to addressing many of these issues have been discussed by Kingsbury and Houser (1998).

Despite the challenges involved in developing adaptive tests for students, several have been developed and are broadly used. Among these are tests developed for modest-stakes, interim assessment of student performance, such as the Scholastic Reading Inventory (Scholastic, 2007), STAR Reading and STAR Math (Renaissance Learning, 2010), and the Measures of Academic Progress (MAP: Northwest Evaluation Association, 2009). Others have been developed to meet the higher-stakes needs of the *No Child Left Behind* legislation (NCLB, 2002) such as the Oregon Assessment of Knowledge and Skill (OAKS: Oregon Department of Education, 2010) and the Delaware Comprehensive Assessment System (DCAS: Delaware Department of Education, 2011).

This study will focus on the MAP system. This system includes approximately 600 operational adaptive tests, including reading, mathematics, language, and science tests designed to align with content standards for each state. It uses measurement scales that were originally developed in the 1980s, and uses thousands of test questions that have been added continually from that time to the present.

All K-12 adaptive achievement tests have required the creation and maintenance of a measurement scale that provides consistent meaning to student scores across years. In each of the operational tests mentioned above, Item Response Theory (IRT: Lord & Novick, 1968; Lord, 1980) has been used to create the measurement scales. IRT allows the development of measurement scales that support normative and criterion-referenced interpretation. It also can provide a scale that may be stable across tests and time, if the assumptions of the model are met.

IRT 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 (Ban, Hanson, Wang, Yi, & Harris, 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, using these items to facilitate calibration of 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.

This study investigates the qualities of stability in the scales that are used for measurement within the MAP system. Specifically, the study examines the extent to which IRT difficulty estimates remain constant over a prolonged period of time. In addition, the study examines whether and to what extent changes in item difficulty

estimates might influence student scores. This study follows from scale stability studies that have been done in the past.

Previous studies of scale stability

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 rather than a set of items used in a single test. Second, it uses measurement scales that have 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 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 Rasch Unit (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

There were 3,091 mathematics items and 1,728 reading items 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 test 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, 2,359 items were available for use in mathematics, and 1,392 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 (2,204 items in mathematics and 1,253 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 0.967 in mathematics and 0.976 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 0.95 to 0.99, and the correlations for reading items ranged from 0.96 to 0.99. 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 do not 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 = 2,359) is nearly twice as large as the sample in reading (N = 1,392). Due to this discrepancy in sample sizes, we would expect to see more extreme differences in observing the mathematics items.

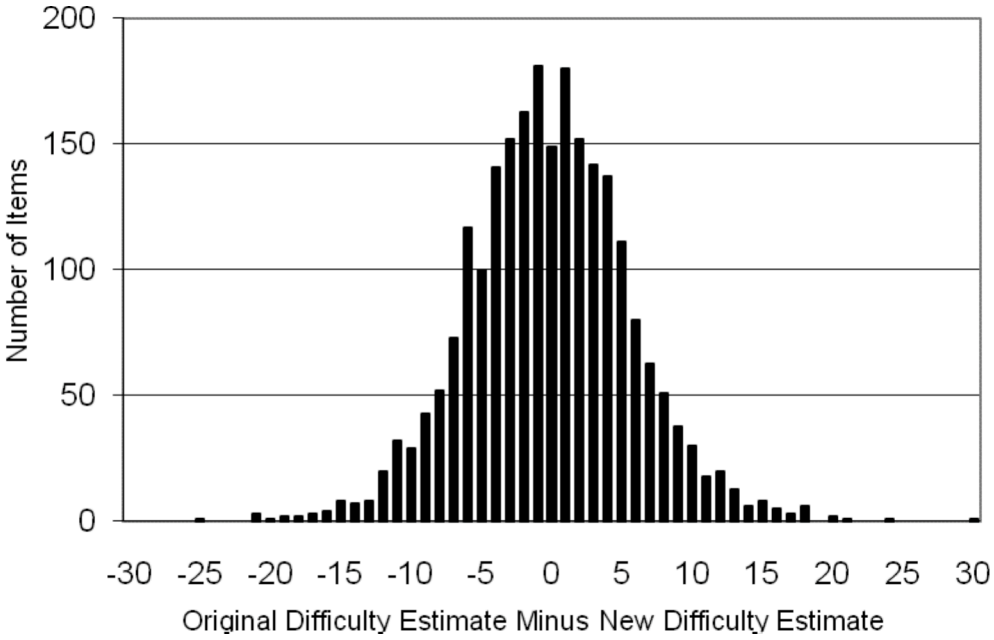

Figure 1. Frequency of mathematics items as a function of the difference between the original item difficulty estimate and the new item difficulty estimate on the RIT scale (rounded to the nearest integer).

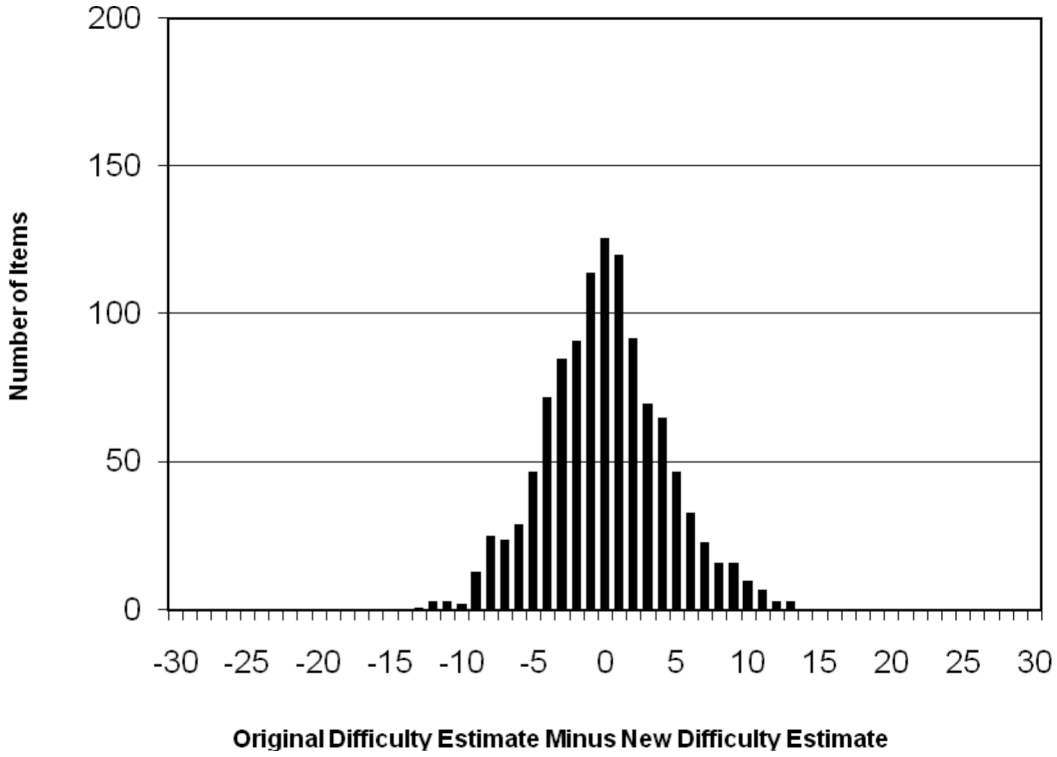

Figure 2. Frequency of reading items as a function of the difference between the original item difficulty estimate and the new item difficulty estimate on the RIT scale (rounded to the nearest integer).

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.

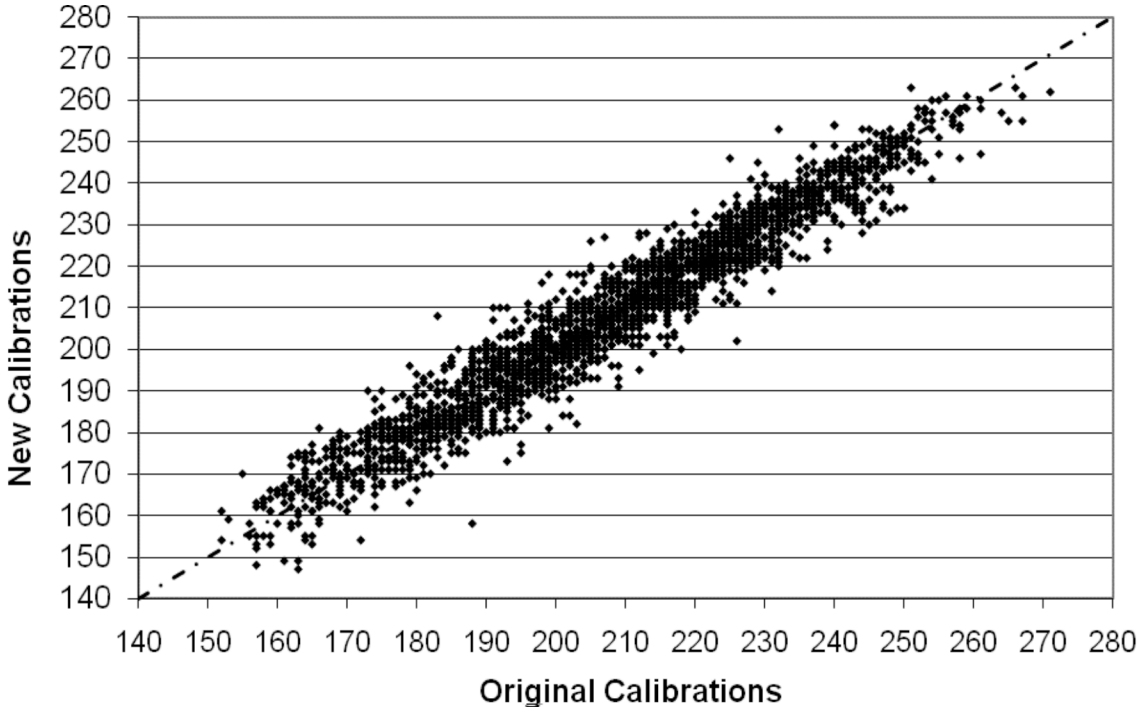

Figure 3. Relationship of original and new item difficulty estimates on the RIT scale for 2,359 items in mathematics with superimposed identity line (r = 0.967).

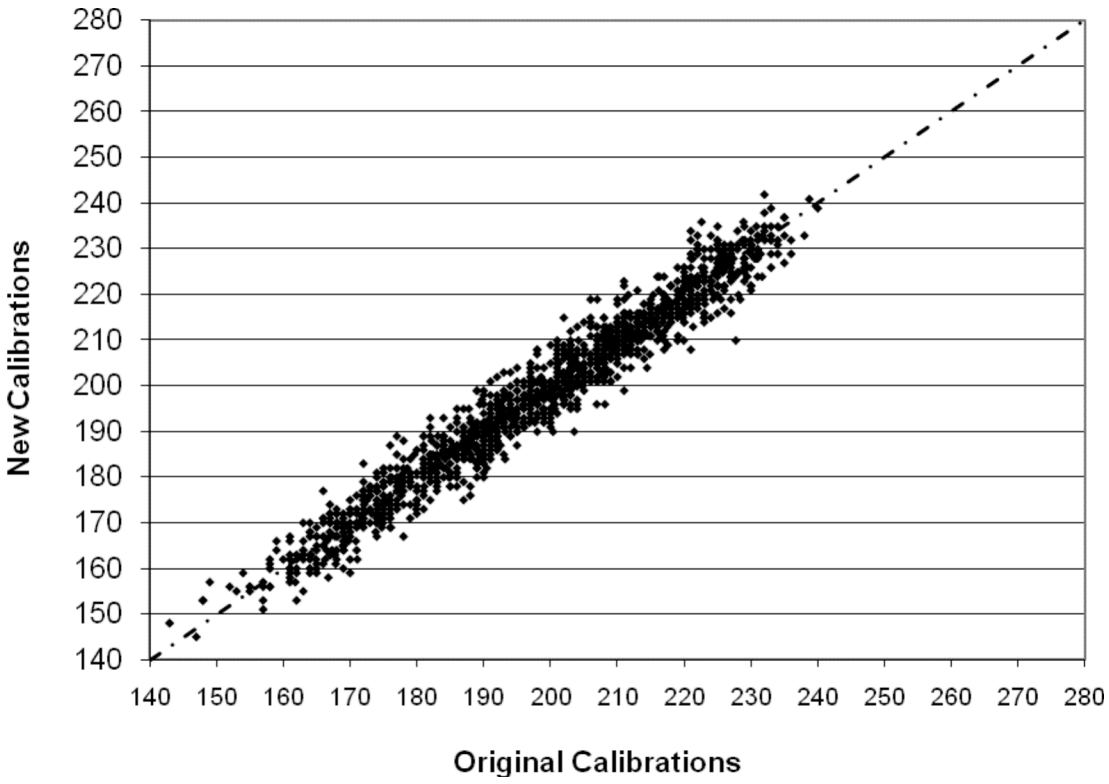

Figure 4. Relationship of original and new item difficulty estimates on the RIT scale for 1,392 items in reading with superimposed identity line (r = 0.976).

The directional drift (bias or average difference) in item difficulty estimates was -0.11 RIT points in reading and -0.17 RIT points in mathematics. To put this difference in context, the standard deviation of students' scores in spring of sixth grade in the most recent norming study done using these measurement scales (NWEA, 2008) was 14.0 RIT points in reading and 15.9 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 0.01 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 3.0 RIT points in reading and 4.0 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 noting. 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.

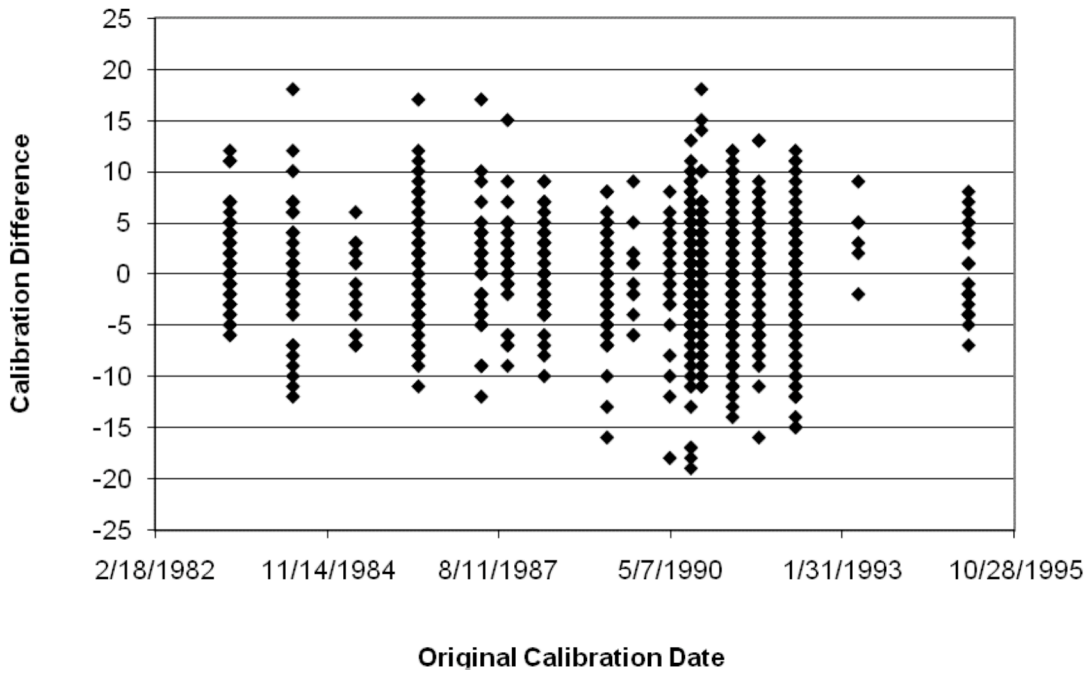

Figure 5. Differences between original RIT difficulty estimates and new RIT difficulty estimates as a function of initial date of calibration in mathematics.

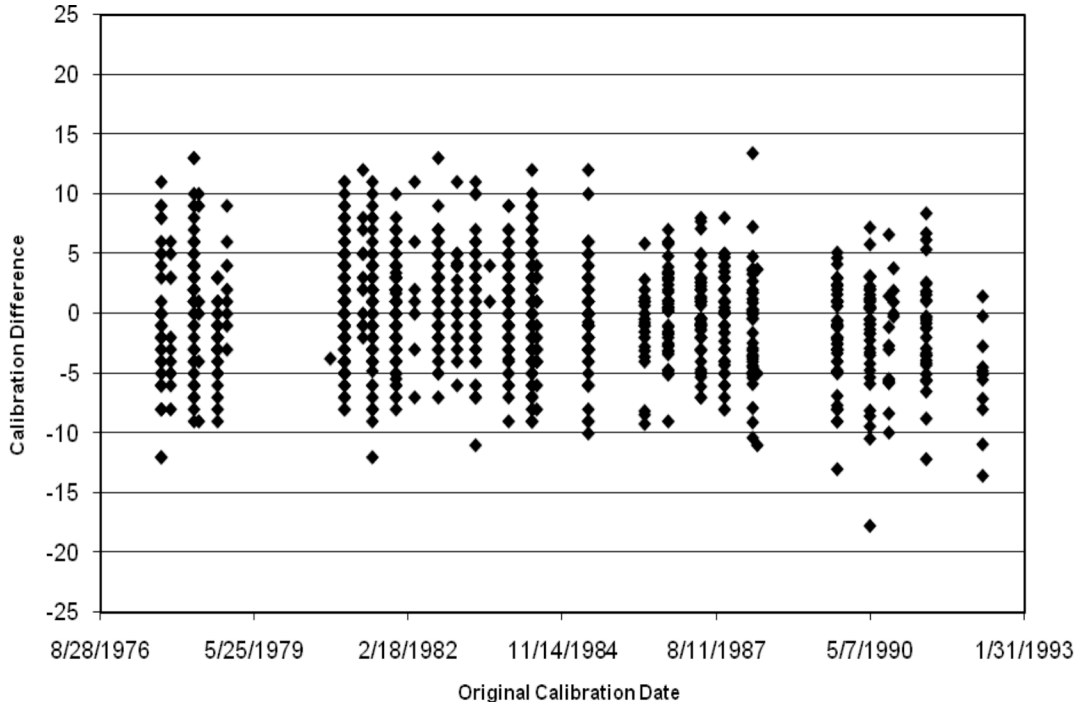

Figure 6. Differences between original RIT difficulty estimates and new RIT difficulty estimates as a function of initial date of calibration in reading.

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 because they are so similar.

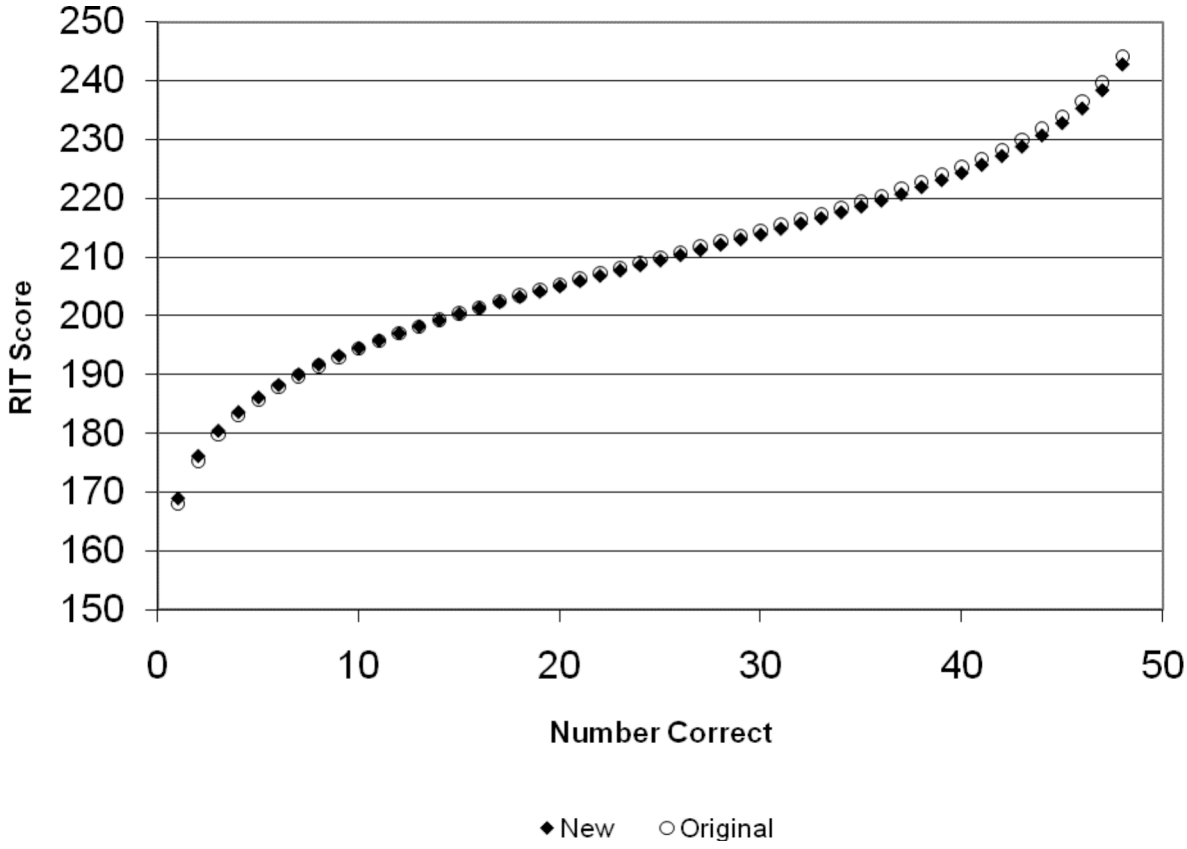

Figure 7. RIT scores as a function of obtained number correct score calculated using original and new mathematics difficulty estimates.

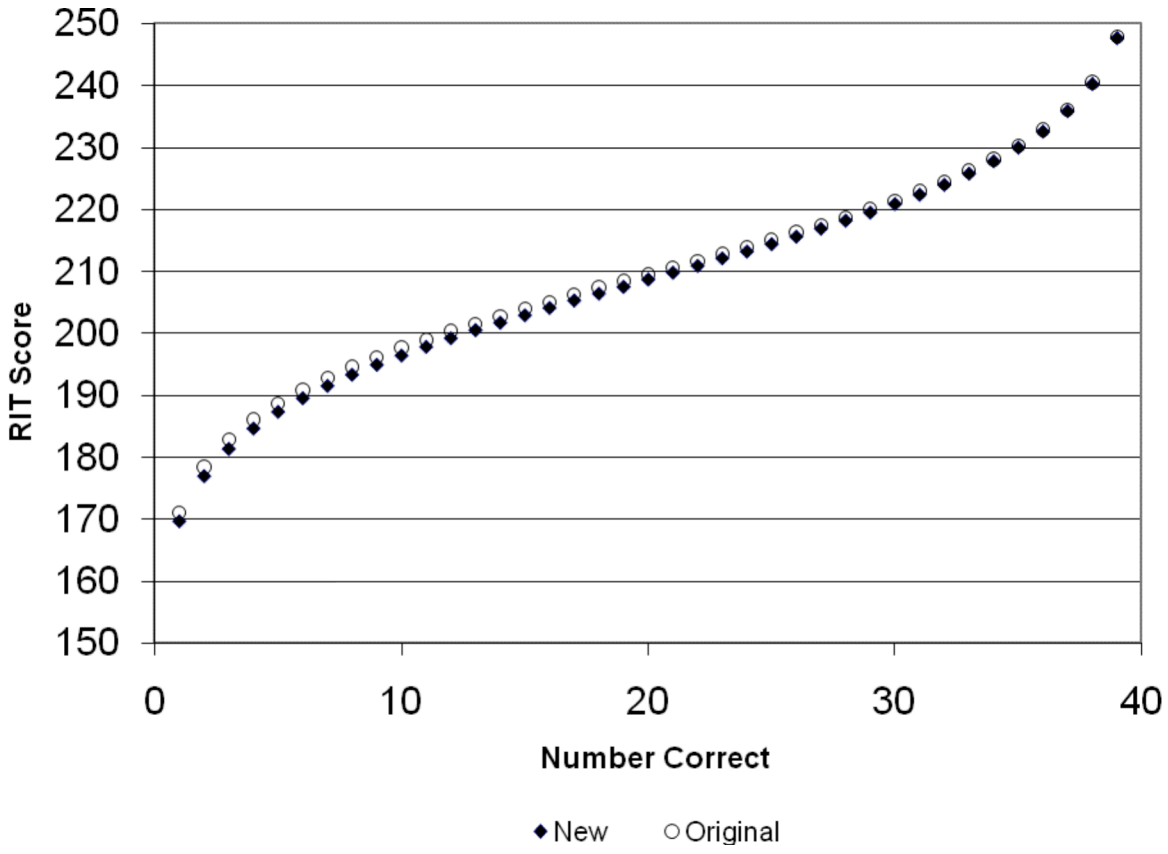

Figure 8. RIT scores as a function of obtained number correct score calculated using original and new reading difficulty estimates.

For the mathematics test, the maximum difference that occurred was 1.1 RIT points and the average magnitude of difference was less than 0.5 RIT points. For the reading test, the maximum observed difference was again 1.1 RIT points, and the average magnitude of difference was 0.7 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 1.0 RIT point in 99 of 100 cases. Since the standard error of a score on one or these tests

would be approximately 4.0 RIT points, the impact of the change in calibrations would not be expected to change instructional decisions.

Discussion and Conclusions

Over the past decade, use of adaptive testing to measure student achievement has exploded. Technology available in the classroom has combined with the need for additional measurement for accountability and student intervention to increase demand. At the same time, operational systems have increased in availability. It seems that the questions that need to be answered in the next decade are shifting from operational capacity to the quality of information that the systems provide and the connections to external referents that the systems provide. This study is a step in the direction of discussing the quality of scores provided from the systems.

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 1.0 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.

Delaware Department of Education (2011). *Delaware Comprehensive Assessment System (DCAS) online test administration manual.* Dover, DE: Author.

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.

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.

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.

No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 U.S.C. § 1425 (2002).

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

Oregon Department of Education (2010). *2009–2010 Technical Report Oregon's Statewide Assessment System Annual Report.* Salem, OR: Author.

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

Renaissance Learning (2010). *The foundation of the STAR Assessments*. Wisconsin Rapids, WI: Author.

Scholastic, Inc (2007). *Scholastic Reading Inventory Technical Guide.* New York, NY: Author.

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.