Growth, Precision, and CAT: An Examination of Gain Score Conditional SEM

Tony D. Thompson

Research Report

December 2008

PEARSON

Using assessment and research to promote learning

Abstract

Monitoring the growth of student learning is a critically important component of modern education. Such growth is typically monitored using gain scores representing differences between two testing occasions, such as prior to and following a year of instruction. The current paper examines the precision of gain scores, and hence their potential meaningfulness, within an item response theory (IRT) framework. The conditional standard error of measurement (CSEM) of the gain score, calculated as a simple function of the component scores from which the gain score is derived, is used to evaluate gain score precision under two different conditions. In the first condition, an example vertical scale is developed using state testing program data to demonstrate that large variation in measurement precision is to be expected when reporting gain scores. Several graphing techniques are utilized to depict visually how measurement precision varies across the distribution of proficiency. In the second condition, adaptive testing is demonstrated to substantially improve the measurement precision of gain scores in comparison to traditional, fixed testing. The potential benefits of adaptive testing for gain scores are discussed, as are the real-world limitations.

Acknowledgements

A version of this manuscript was presented at the 2008 Annual Meeting of the National Council on Measurement in Education in New York. The author wishes to thank Denny Way and Ye Tong for their helpful comments on an earlier version of this paper. The author also expresses thanks to Bob Dolan, who made a number of suggestions that substantially improved the paper.

Growth, Precision, and CAT: An Examination of Gain Score Conditional SEM

Measurement of student growth in learning is an important topic for K-12 state testing programs, both in terms of school accountability as well as for reporting progress of individual students. Recently, Ho (2007) provided a brief status report of growth models in the field of educational measurement, citing a surging interest in growth modeling that is likely to continue given draft reauthorization proposals for the No Child Left Behind Act (NCLB) that prominently feature growth models. Despite this surge, however, relatively less attention has been given to whether reported measures of growth are precise enough to be meaningful and useful. For example, it has been acknowledged that tests built for school accountability under the NCLB status model are unlikely to provide ideal measures under a growth model (e.g., Steering Committee of the Delaware Statewide Academic Growth Assessment Pilot, 2007). Most non-adaptive proficiency tests (i.e., those consisting of fixed, predetermined content), such as those designed to assess students' knowledge, skills, and abilities vis-à-vis grade-level academic standards, measure the central part of the proficiency distribution much more precisely than the extremes of the distribution. In many cases, the non-central regions will be measured poorly. As a consequence, growth scores for students scoring above- and below- average will be less precise than for students whose scores are near average. Furthermore, when selecting a measure of student growth it is important to consider the inherent precision of the measure. A common way to represent growth for a particular student is to simply take the difference between the student's current year score and their previous year's score. Assuming the two scores are on the same scale, the difference can be thought of as the amount the student has "gained." Although policy makers find gain scores appealing because the scores are easily understood by stakeholder groups, psychometricians have long questioned the reliability of gain scores. The issue of score precision is important because even the most elaborately designed vertical scale or growth model will be useless if the scores reported from it lack precision.

In this paper, the precision of the most basic of growth measures, the simple gain score, is examined within an item response theory (IRT) framework. Because comparatively little of the vast psychometric literature on gain scores has approached the subject from an IRT perspective, some background on IRT gain score precision is provided. To this end, we explain that the conditional standard of error of measurement (CSEM) of the gain score is calculated as a simple function of the individual CSEMs of the two component scores from which the gain score is derived. Next, the variation in measurement precision that can be anticipated when reporting gain scores is demonstrated by developing an example vertical scale (for equating test scores across multiple grade levels) using item parameter estimates from a state testing program. This variation, we argue, likely renders non-informative any vertical scales developed from conventional (non-adaptive) tests due to lack of score precision. Finally, in another demonstration using item parameters from a state testing program, gain scores from a computer adaptive test (CAT) are compared to those from a non-adaptive version. Discussion of the potential benefits of CAT and the likely limits of those benefits is provided. Although the main context for the paper is vertical scales in K-12 assessments, the methods employed to investigate gain scores apply to any situation where IRT-based gain scores are used.

Study Design

Gain Scores

Scores calculated by subtracting the test results of the same student on two separate occasions has been called by several names in the literature, including change, difference, deviation, gain, growth, or progress scores. The classic pretest-posttest experimental design is a common example of a difference score; in this design, the same measure is taken of subjects before and after a treatment condition, with the difference between measures taken as an indicator of treatment effect.

Difference measures are commonly used in testing. In the K-12 state testing arena, an important difference measure is the growth or progress a student exhibits from one year to the next. The simplest growth comparison for an individual student is to compare current year's scores with previous year's scores for particular subject areas. If a common scale can be established between grade levels with a vertical scale, a student's scores from two different years can be directly compared. The difference between these scores on the vertical scale represents the demonstrated change in student proficiency on the measured constructs and is often referred to as the student's growth or progress score. In this paper, the term gain score is adopted.

As described by a number of authors (e.g., Singer and Willett, 2003, p.10; Willett, 1997), simple difference or gain scores are limited as tools for studying change. A score based on the difference of just two measures, even two reasonably reliable measures such as successive end-of-year NCLB tests, is unlikely to fully and precisely measure the individual growth experienced by each student. Although end-of-year tests tend to be quite broad in their content coverage, it is unrealistic to expect a single test given on a single occasion to fully measure the knowledge, skills, and abilities gained by a student in a given school year. Also, while end-of-year tests are typically reliable enough to provide a reasonable rank ordering of students, the difference of two measures is generally less reliable than the measures themselves, as discussed below. More powerful procedures for measuring change can be employed when data are collected from three or more points in time. One advantage gain scores do have, however, is the ease with which they may be explained to lay audiences. Because of the intrinsic simplicity of gain scores, they may be advocated for policy reasons and thus it is important to examine the limits of gain score measurement precision. If nothing else, such an examination can help inform when the gain score will be too imprecise to be useful.

The inherent reliability of gain scores has long been a controversial topic in psychometrics. Most of the literature examining this issue has focused on the classical test theory representation of gain scores, namely

$$S_1 = T_1 + E_1, \tag{1}$$

$$S_2 = T_2 + E_2, \tag{2}$$

$$G = S_2 - S_1 = T_2 - T_1 + E_2 - E_1, \tag{3}$$

where S, T, and E designate the observed, true, and error scores respectively, G represents the gain score, and the subscripts refer to either the first or second testing occasion. As shown in Equation 3, both true and error components are subtracted to determine gain score variance. While subtracting true scores generally diminishes true score variance, subtracting uncorrelated error scores adds to error score variance. This combination of effects tends to greatly diminish reliability (but see Williams and Zimmerman, 1996, who show this result is not inevitable).

The current paper presents the point of view espoused by Mellenbergh (1999) and Fischer (2003) that while reliability can be a useful measure, it is a relative rather than absolute indicator of a test’s measurement precision. By definition, reliability is the ratio of true score variance to total variance, and can be represented as a ratio of true variance to error variance. For example, Table 1 gives the ratio of true to error variance (sometimes called the signal-to-noise ratio) for selected reliabilities. Reliability does not, however, indicate the absolute value of the error variance, which could potentially be low even when the reliability for a certain population is low. For example, a particular assessment given to two different populations might result in equal error score variances for the two groups (i.e., the reliability denominator, or *noise*), but the reliability of the scores for the two groups could be markedly different due to true score variance differences between the groups (i.e., the reliability numerator, or *signal)*.

Table 1.
Example Reliability and True Score Error Variance Ratio Values

Reliability	Ratio of true to error variance
.9	9:1
.8	4:1
.7	7:3
.6	3:2

In contrast to the classical test-derived reliability measure, the IRT conditional standard error of measurement (CSEM) provides an absolute measure of test precision for a given score scale. It has the further advantage of varying as a function of student proficiency, rather than just being a single value across all proficiencies. Since test precision can thus be better represented across a range of student proficiencies, it is a more appropriate indicator for tests that follow an IRT model.

Despite the potential advantages of obtaining an absolute measure of precision for a score scale, only a limited number of studies have taken IRT approaches to studying gain scores. (For an excellent review of the change literature, both for classical test theory and for IRT, the reader is referred to Wang and Wu, 2004). Of the few IRT studies that have examined change, most have focused on the Rasch model (e.g., Fischer, 2003; Wang and Wu, 2004). The current paper, however, focuses on the three parameter logistic (3PL) model (Birmbaum, 1968) and its polytomous generalizations (for review, see Ostini and Nering, 2005). May and Nicewander (1998) used the 3PL model to examine a gain score problem they called scale distortion, which stems from having a pretest that is too easy and thus induces a ceiling effect. They showed that IRT scoring, possibly combined with adaptive testing, could reduce scale distortion for gain scores. One important, relevant implication of their study is that the vertical scale must be well-formed and appropriate for the application; a poorly defined vertical scale will likely produce poor gain scores.

Using a different approach, Nicewander (1991) proposed a modified gain score to increase item reliability for pretest/posttest gain scores. The modification, however, is not relevant to vertical scales, as it only applies to situations where the pretest and posttest are the same. For the traditional gain score, the study found item reliabilities to be extremely small except for the case of highly discriminating items accompanied with a large change in proficiency. May and Jackson (2005) based their approach on that of Nicewander (1991) and explored item level reliabilities for the 3PL model. In general, they found similar results, with very small item reliabilities for items with typical or low discrimination (< .05 reliability for *a*-parameter values less than 1.5). Their results were taken as further evidence of the inherent unreliability of gain scores. Both the Nicewander (1991) and the May and Jackson (2005) papers highlight the potential lack of gain score precision that may occur for low discriminating items. However, rather than focusing on item reliability as these studies did, the current study argues that using the IRT CSEM is a more effective and powerful tool for studying gain score precision.

Gain Score CSEM

In this section of the paper, the gain score CSEM is derived. Following the derivation is a demonstration of how the gain score CSEM might look when IRT is the model underlying the vertical scale upon which gain scores are based . The context for this demonstration is an IRT vertical scale for a state testing program that links adjacent grades from 3-8.

The CSEM of gain scores follows the same definition as the CSEM for any score: CSEM is the square root of the conditional error variance of the gain score. However, unlike the CSEM from a single test score, for a gain score there are two true proficiencies to condition on, the previous grade's theta and the current grade's theta. As an example, consider the case of gain score G, defined as the difference between two scores from two occasions but scaled to a common metric. Hence,

$$G = S_2 - S_1, \tag{4}$$

where the subscript refers to either the first or second occasion. For a given pair of theta values on the IRT vertical scale, θ_1 and θ_2, assume that the error from occasion one is uncorrelated with the error from occasion two. The perspective given here is that θ_1 represents a student's proficiency in year one and θ_2 represents that same student's proficiency in year two. The conditional error variance of the gain score is then given by,

$$Var(G \mid \theta_1, \theta_2) = Var(S_1 \mid \theta_1) + Var(S_2 \mid \theta_2) \tag{5}$$

The gain score conditional error variance is the sum of the conditional error variances of the individual scores. The CSEM of the gain score is the square root of the conditional error variance. Therefore, the CSEM of the gain score can be calculated from the CSEM of the two individual scores as given by,

$$G(\theta_1, \theta_2)_{CSEM} = \sqrt{[S_1(\theta_1)_{CSEM}]^2 + [S_2(\theta_2)_{CSEM}]^2}\ . \tag{6}$$

A relevant point taken from Equation 6 is that the CSEM for the gain score must be larger than the CSEM from either of the two component scores (assuming these are both non-zero). In this sense, gain scores must be less precise than the scores that they are derived from.

Rather than conditioning on the true values from the two component tests, it might seem simpler to condition on the true gain. That is, let

$$\eta_i = \theta_{2i} - \theta_{1i}, \tag{7}$$

where η_i is the true gain for student *i*. A student’s expected gain score equals their true gain. That is, in a hypothetical experiment where a student takes the test many times but without the influence of practice or learning effects, the average over replications equals the student’s true gain. However, the standard deviation across replications for the student will not necessarily equal the replication standard deviation for other students with the same true gain. Because students with the same true gain may have proficiencies at different points of the theta distributions, conditioning must occur at the level of the individual test, rather than on a student’s true gain.

Gain Score CSEM Demonstration

Thompson (2007) demonstrated how the CSEM of an IRT growth score might work in practice using data from two large-scale reading comprehension tests from a state testing program. A portion of those results is duplicated here. The two reading comprehension tests used were from grades three and four from a state-wide administration in 2006. The tests were mostly comprised of multiple choice items, but also included two three-point constructed response items. Summary information about these tests is given in Table 2.

Table 2.

Summary Information of Tests Studied

Test	Total points	Average IRT a-value	Average IRT b-value	Average IRT c-value
Reading grade 3	44	1.09	-0.75	0.19
Reading grade 4	46	0.83	-1.02	0.13

The two grades were linked through a set of common items that were administered to both grades. For clarity of exposition, the actual vertical scales developed for the state in question are not used here. Instead, a slightly simplified version is utilized to serve as an example of what would likely be found in practice. The vertical scale is the grade 4 theta scale, with the grade 3 theta scores transformed to the grade 4 metric. The grade 3 theta was transformed as follows,

$$\theta_{New} = \theta_{Old} - .4 \,. \tag{8}$$

It is common for a reported vertical scale to be a linear transformation of the theta scale. Because a linear transformation of a scale equally applies to the CSEM, the theta scale is used here as the reporting scale. A more complicated method of deriving the reporting scale would probably have little effect on the results. Once the IRT CSEM for each grade was found, the CSEM for the gain score was found using Equation 6.

Figure 1 presents a 3-dimensional graph of the CSEM for the theta metric gain score for the reading tests. The two lower axes represent true theta for the two grades after grade 3 was transformed to the grade 4 scale. The theta scales are plotted from -2 to +2. The vertical axis on the plot is the CSEM of the gain score on the common grade 4 theta scale. The perspective given is looking down from approximately 45º above the graph. The figure shows that the lowest CSEM values are associated with theta values for the two grades in a small region around -1.2 to -.6; CSEM values in this range are approximately .4. For theta values between -2 and approximately .2 for both grades, but outside the previously described region, CSEM values ranged from around .4 to .6. As theta values increased into the positive range for both grades, the CSEM increased as well. The maximum CSEM values were found when both grade three and grade four thetas were greater than 1.4; the CSEM values for this region were approximately 1.4.

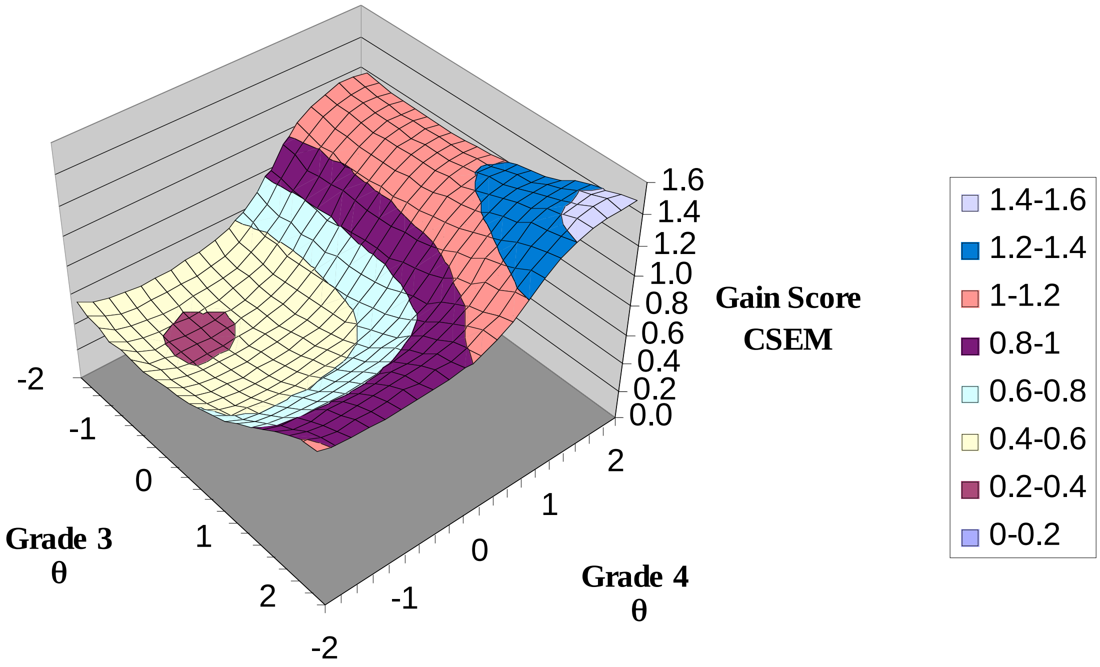

Figure 1. Gain score CSEM plotted for pairs of grade 3 and grade 4 reading proficiency.

The surface in Figure 1 results from both grades' tests having maximal measurement precision between theta values of approximately -1.2 to -.6. Both were relatively easy tests for their respective populations (as can be seen by the low b-values in Table 2). Because the tests worked best for this region of the theta scale, the gain score is also most precise in this region. The further away from this region on the graph one goes, the larger the gain score CSEM becomes. For tests more centrally targeted, the lowest gain score CSEM values would be closer to the center of the distribution.

A three-dimensional plot such as the one in Figure 1, offers the measurement specialist an excellent tool for visualizing gain score precision, especially if the plot is given in grayscale or color to highlight CSEM value ranges. To create the CSEM values for each of the desired pairs of theta points, only the item parameters for the linked assessments are required. Because widely-available software, such as Microsoft Excel™, can be used to create such 3-D plots, special graphing software is unnecessary.

Another way to visualize the gain score CSEM is shown in Figure 2. In this plot, gain score CSEM is plotted against true gain score on the theta scale. As stated before, the lower grade's theta was transformed to the scale of the upper grade before calculating the gain score. As can be seen, the CSEM is not constant for a given true gain score, because the same true gain score can be obtained from multiple pairings of the two grades' proficiencies. For example, a true gain score of zero represents no growth from the lower grade to the upper grade, but it does not specify from which theta value the student failed to grow.

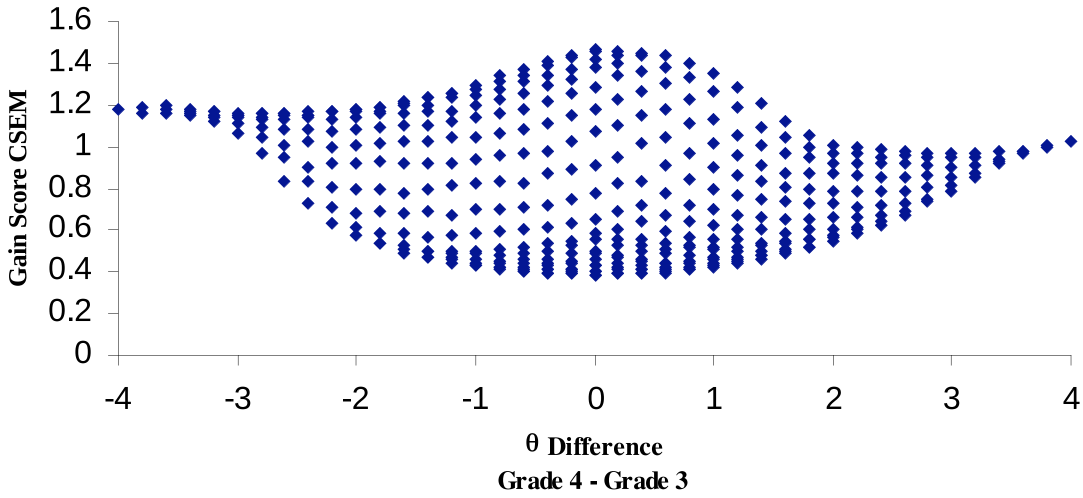

Figure 2. CSEM plotted versus gain score for reading tests.

The other striking aspect of the graphs is that the minimum CSEM is around .4 (corresponding to the small area of points in Figure 1 with gain scores CSEM in the .2-.4 region). Thus if a two-standard deviation confidence interval is used with the CSEM, even in the best case the gain score interval will vary from -.8 to +.8. On the theta scale, this is a quite large range of values. Although this scale was created for demonstration purposes, the implication is that gain score confidence intervals created from these tests are likely to be too large to give precise estimates of gain scores. Since the tests in question are fairly reliable measures of on-grade performance, the further implication is that gain score confidence intervals may be large for *many* educational tests. Of course, these conclusions depend upon the nature of growth that is observed. If the variability of observed growth is large, then reasonably valid comparisons between individuals may still be possible. Also, if an individual's observed growth is much larger than the associated CSEM, then we may be confident that the individual did experience growth in learning.

CAT Gain Score Precision

The overall conclusion from the gain score study was that a vertical scale developed from a typical K-12 testing program might support reporting of informative gain scores for some students, but that for a large proportion of students gain scores would be non-informative (Thompson 2007). This finding is a result of individual grade-level tests not being designed to measure all students with equal precision. Equation 6 shows that the gain score measurement precision can be no better than the precision of either of the two components that contribute to the score; thus, a component score of low precision necessarily results in gain score of low precision. One natural way to address the issue of obtaining high measurement precision across the proficiency distribution is through adaptive testing (Van der Linden & Glas, 2000; Wainer, 2000).. Because adaptive testing can increase measurement precision compared to a conventional test, especially in proficiency regions where linear tests tend to yield little information, a computer adaptive test (CAT) design may enable reliable growth measurement for all students.

A recent simulation study by Kang and Weiss (2007) explored the use of adaptive testing to study individual change. In their study using the 3PL model, simulated examinees were administered either a 50-item CAT or 50-item conventional test at two points in time. They simulated three levels of average growth: .50 (low), 1.0 (medium), and 1.5 (high) theta units. They found that the conventional test detected significant change (i.e., non-overlapping error bands between two testing occasions) best for the proficiency levels that the test was targeted to. The CAT, however, detected significant change equally well across the proficiency scale. In addition, the CAT was superior to the conventional approaches in measuring change, in terms of correlation with true change, root mean square error, and bias. The results of the study strongly supported use of adaptive testing to measure individual change, especially in the medium and high growth conditions. Also, the results indicated that gain scores from conventional tests are generally not useful for measuring change.

Two limitations of the Kang and Weiss study limit its applicability to operational large-scale testing. First, the item selection algorithm in the study's CAT simulation did not control for either content or item exposure. Second, the IRT parameters of the items in the pool were randomly generated from an ad hoc distribution, rather than being based on empirical data from an existing testing program. To investigate the potential of CAT in the context of a vertical scale to obtain precise gain scores across the proficiency distribution, and thus effectively measure individual change under real-world constraints, a second study was conducted. A second objective of this study was to further highlight how the CSEM function from Equation 6 is useful for evaluating the precision of IRT-based gain scores.

CAT Simulation Method

A computer simulation method was used to compare the gain score precision from traditional ("paper") and adaptive versions of a mathematics test. A simulation of a vertical scaled adaptive NCLB test was not attempted; rather, data from an existing CAT simulation were used. Thompson and Way (2007) created a detailed simulation of a state mathematics graduation test to compare several different CAT designs and resulting score comparability with a paper version. Although the purpose of that study was different and no vertical scale was simulated, these data permitted a realistic CAT simulation consisting of a realistic CSEM function for CAT and paper test comparison, rather than use of mocked-up item parameters and hypothetical content constraints. These data provided the "upper" grade CSEM for purposes of the current study. For the lower grade CSEM, a few assumptions were made. First, as for the grades 3 and 4 reading tests described previously, the tests from two adjacent grade levels were assumed to similarly target their respective populations. Second, CAT item pools for the two grades were likewise assumed to be of similar breadth and depth for their respective populations. Finally, it was assumed that the hypothetical lower and upper grade tests measured the same constructs, but that the theta scale of the lower grade was .40 points lower than the upper grade scale. The .40 value was chosen because it matched the empirical difference found between the two reading tests described earlier and was typical of adjacent grade differences for the vertical scale studied in Thompson (2007).

Thus, the CSEM functions for both the upper and lower grade tests were assumed to be identical, except for the adjustment to put the lower grade theta on the upper grade scale. The effect of this was that the lower grade paper test measured the lower end of the proficiency scale better and the upper grade paper test was superior at the upper end of the scale. Note that a zero theta adjustment would simulate a pre/post test scenario. While the actual growth of students going from the lower to the upper grade was not modeled, the assumed scale difference implies a lower grade student would need to gain .40 units on the vertical scale to maintain their relative standing in proficiency. Note that this level of average growth corresponds closely to the low growth condition of the Kang and Weiss (2007) study.

Relevant details of the procedures and algorithms implemented in the study are described below. Alternate algorithms could be considered in future work, but as the study is exploratory in nature, conventional, realistic methods were chosen.

Data and Item Parameters

Simulations were based on data from a statewide grade 11 mathematics test administered in spring 2003. The 60-item operational test consisted of discrete four-option multiple-choice items and a small number of grid-in items (about 9%). There were 60 different sets of 10 field test items embedded in different versions of the test.

The initial item pool for the CAT simulations was comprised of the field-test items from the paper, a total 600 of items. The 60 operational questions comprised the conventional test form to which the CAT results were compared. Table 3 provides the numbers of items in each content objective for the operational test and CAT item pool.

Table 3.

Numbers of Mathematics Items by Objective Areas

Mathematics test objective	# items in paper test	# items in CAT pool
Objective 1	5	47
Objective 2	5	59
Objective 3	5	61
Objective 4	5	61
Objective 5	5	48

Objective 6	7	61
Objective 7	7	71
Objective 8	7	81
Objective 9	5	48
Objective 10	9	63
Total number of items	60	600

3PL calibrations, carried out using BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1999), were conducted on the item pool and served as both the true parameters and as the parameter estimates by the CAT and paper test simulations. That is, estimation error of the model parameters was not considered in the simulation.

Item Selection Algorithm

The CAT used a fixed length test length of 35 items using maximum information item selection control. This value was chosen in the Thompson and Way (2007) study as the test length that allowed the CAT to match or exceed the measurement precision of the comparable paper test.

Content Balancing Method. Content was balanced for the 10 objective score areas described in Table 3. The goal was for each objective to have the same proportionally representation in the CAT as in the paper test. The algorithm selected only items from the “most needy” objective at each point in the CAT (ties resolved randomly). Objectives were deemed “most needy” when their proportional representation was most dissimilar to the paper test content distribution.

Theta Estimation. The base ability estimation method used was maximum likelihood. Until at least one incorrect and one correct response occurred, theta was estimated through a step size value procedure. The initial theta was set at -1.0, with theta moving by +1.0 after each correct response or by -1.0 after each incorrect response until maximum (+4.0) or minimum (-4.0) thetas were reached.

Exposure Control Algorithm. The Sympson-Hetter exposure control procedure was implemented (Sympson & Hetter, 1985). The maximum desired item administration rate was set to .15. The calibration of exposure parameters was performed for 20 cycles on samples of 4000/cycle. The thetas used to generate the response data for each cycle were generated from a N(0,1) distribution.

Other Simulation. Details. Simulated response vectors using 41 true proficiency values from -4 to +4 were randomly generated. At each proficiency level, 200 simulated examinees were generated. A total of three replications was performed.

Results

The same pattern of results was found in each of the three replications. For the purpose of simplifying the presentation of results, the tables and graphs below report the first of the three replications performed as they are representative of the other two. Thompson and Way (2007) report on the measurement quality of the 35-item CAT test as compared to the 60-item paper test. The reader is referred to that paper for complete details. Briefly, the CAT version was shown to have the higher correlation with true theta, higher classification accuracy, less biased score estimates, and lower CSEM values compared to the paper version. The CAT also met all content constraints, with all item administration rates less than .2.

Because the gain score CSEM is a function of the lower and upper grade CSEM functions, it is informative to compare the CSEM functions for the CAT and paper versions. These functions are presented in Figure 3, with the lower grade CSEM functions given on the upper grade scale. The CSEM values for the CAT versions are lower in the extremes, but in the center of the scale (from about -.5 to +.5), the CAT and paper results converge. For both the CAT and the paper test, the lower grade test measures more precisely in the lower end of the distribution and less well at the upper end. The converse is true for the upper grade tests. These effects are much less pronounced for the CAT, however.

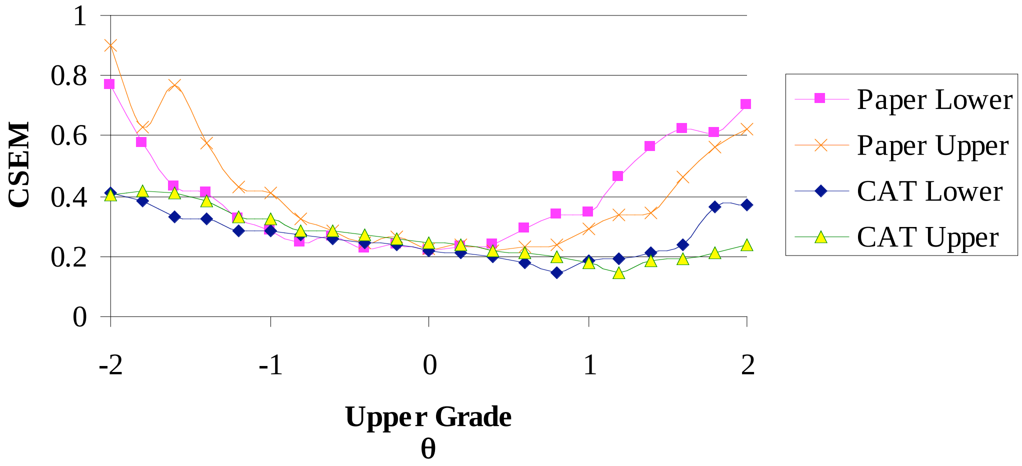

Figure 3. CSEM for CAT and paper test.

The gain score CSEM functions are presented in Figure 4 (paper version) and Figure 5 (CAT version). The surface for the paper test version is similar to that for the vertical gain score from the reading test given in Figure 1; namely, there is a wide range of CSEM values depending upon students' starting proficiencies in the lower grade and their ending proficiencies in the upper grade. Unlike the reading test, however, the math test item parameter values seem well targeted to the population; this is reflected in the wide center region of the plot that shows the two tests are measuring best in middle of the proficiency distribution. Only at the more extreme values of theta does the CSEM function become large.

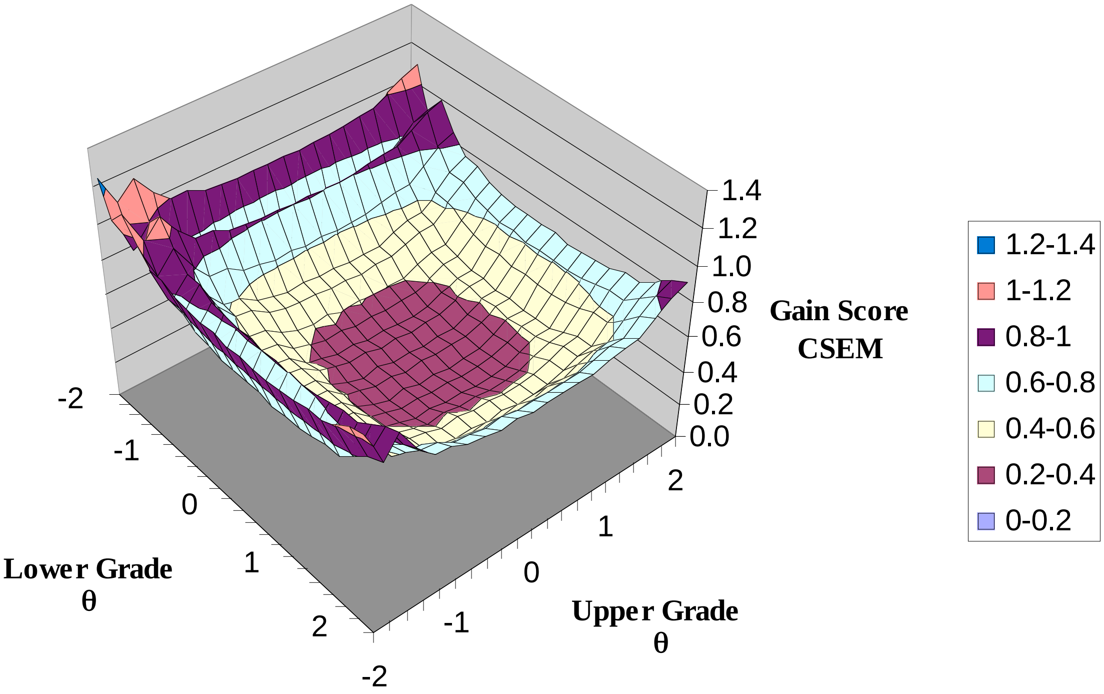

Figure 4. Gain score CSEM paper test.

In contrast, for the CAT version (Figure 5), the gain score CSEM is much flatter. Most of the theta region has CSEM values less than .4, and no values exceed .6.

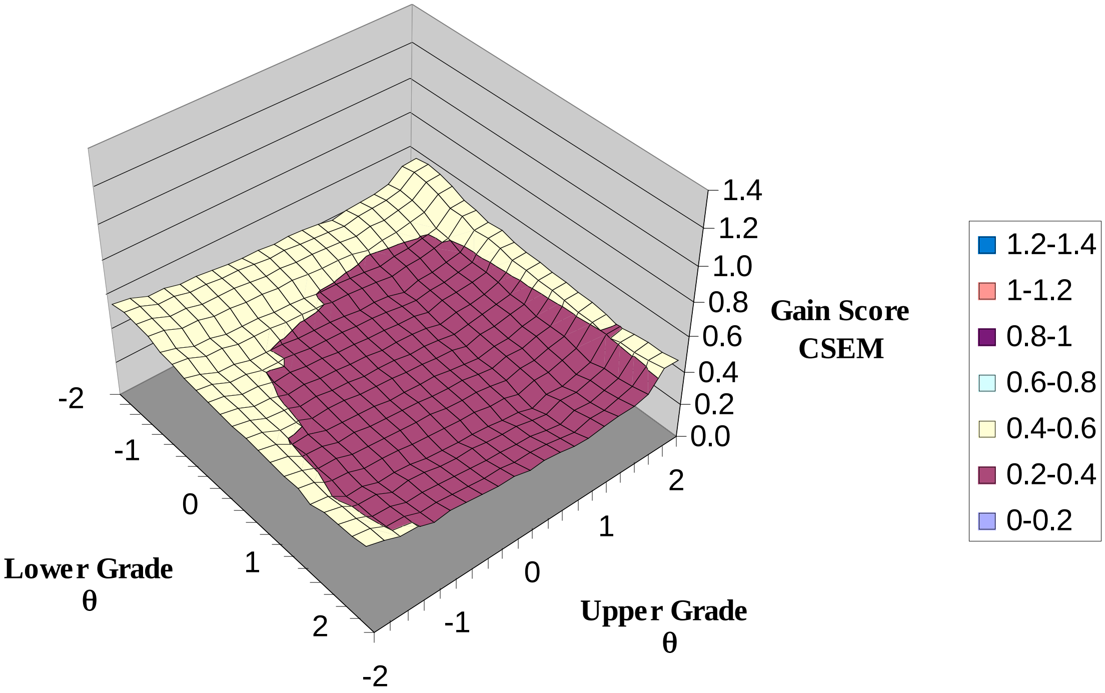

Figure 5. Gain score CSEM CAT test

Figure 6 and Figure 7 present the gain score CSEM plotted against the true gain. In these plots, the lower end is not relevant, as we do not expect many students to have a substantial negative gain; we expect the vast majority of student gain scores to fall in the 0 to 2 region. Although the CAT version results in a much narrower band of CSEM values than the paper version, in the 0 to 2 region of interest there is still a fair degree of variability for the CAT.

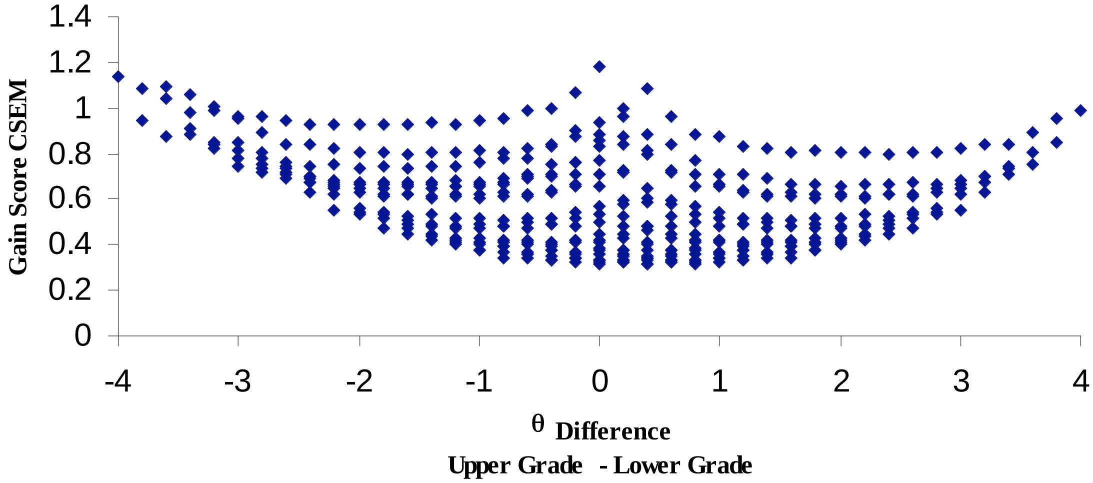

Figure 6. CSEM versus true gain paper test.

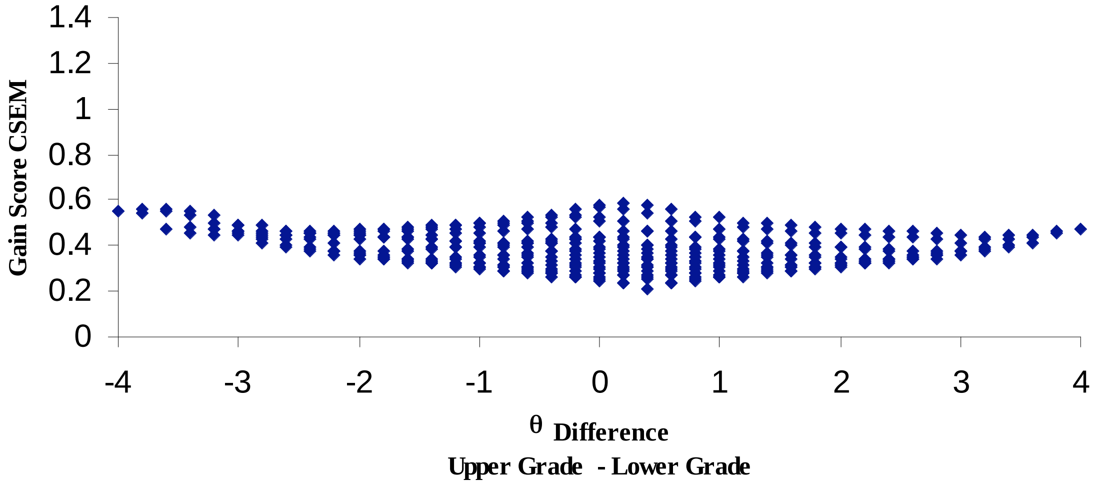

Figure 7. CSEM versus true gain CAT test.

To investigate the variability in CSEM values more deeply, +/-2 confidence intervals were formed for the gain score of .60. The choice of the .60 value was arbitrary, but illustrates an expected gain for a typical student on a vertical scale (note, it is slightly larger than the .40 scale difference found by Thompson (2007) for the reading tests reported earlier in the paper). The intervals are plotted for the paper test in Figure 8 and for the CAT in Figure 9. Note that there are several intervals plotted, because the appropriate CSEM to use in the interval depends upon the student's lower and upper grade scores.

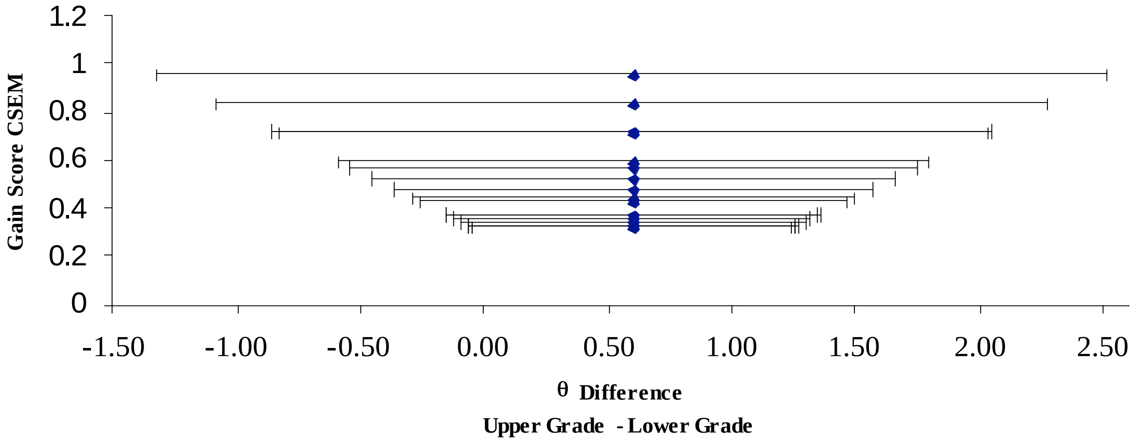

Figure 8. +/- 2 CSEM confidence intervals for .6 gain paper test.

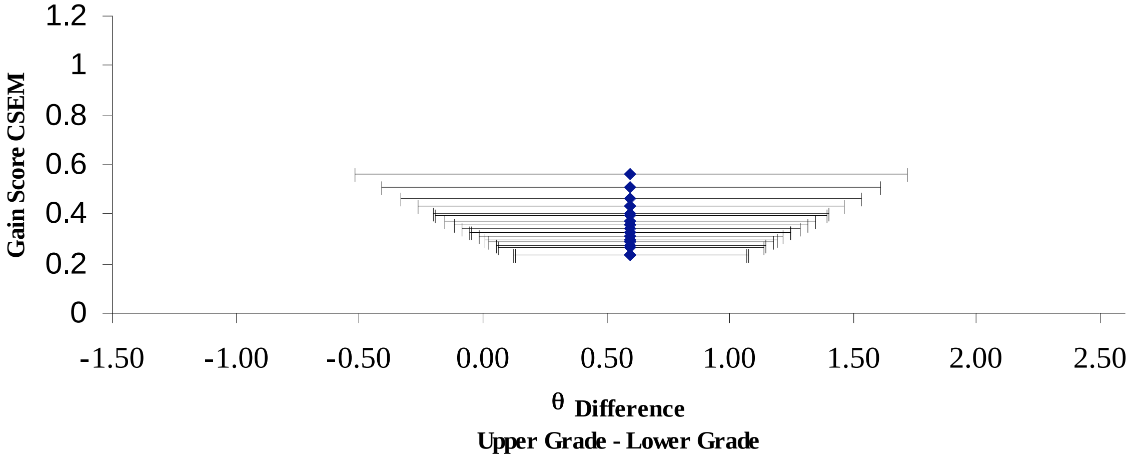

Figure 9. +/- 2 CSEM confidence intervals for 6 gain CAT test.

Consistent with the other results, the paper test gives similar confidence intervals to the CAT for some of the CSEM values, but the intervals are dramatically wider for other CSEM values. For the .6 gain score, the smallest CAT confidence interval is .13 to 1.07 and the largest -.52 to 1.72. This compares to the smallest paper test confidence interval of -.04 to 1.24 and the largest of -1.32 to 2.52. In general the widths of the paper test confidence intervals compromise the meaningfulness of the gain score. The CAT test has much more consistent interval lengths, though there is still some variability. The CAT intervals, as a whole, are also much smaller than the paper test intervals, but whether the intervals are small enough to make the gain scores informative is somewhat questionable.

Discussion

In this paper, the precision of the simple gain score derived from an IRT vertical scale was examined. It was shown that the CSEM for the gain score is a straightforward function of the two CSEM functions for the lower and upper grade measures from which the gain is calculated, and that the gain score inherently must be less precise than the lower and upper grade measures. Furthermore, CSEM values can vary for individuals with the same true gain, and for conventional tests, CSEM values will vary substantially as a function of eccentricity from average score.

Although the study was exploratory in nature, several broad conclusions are supported by the results. The primary point of the paper is that however a growth measure or vertical scale is formed, the precision of the resulting scores must be considered. The measurement precision of scores based on vertical scales has not been studied sufficiently. Part of the process for determining testing program growth score models must be an evaluation of whether reported scores will be precise enough to support meaningful decision-making. This point is underscored by the often poor precision of the paper test gain scores. These results agree with much of the long history of research on the reliability of difference scores. However, much of the previous research focuses on the global and population-dependent reliability measure rather than a conditional measure. By using the CSEM, it is clear that measurement precision of gain scores can vary greatly within a given pair of tests, with differential impact on students. For example, if gains are large relative to the error of measurement, then gains scores are meaningful even if the error is fairly large in an absolute sense. Studies comparing typical observed gains for an operational vertical scale to the CSEM would provide well-needed empirical evidence for the meaningfulness of gain scores. For conventional paper tests, however, the observed gain scores are likely to be relatively small compared to the error of measurement.

Adaptive testing was examined as a possible method of improving the CSEM of gain scores to an acceptable level. Here, the findings from the small simulation study were mixed. The CAT yielded CSEM values that were both much smaller and more consistent in magnitude than for the paper version of the test. However, it was unclear from the simulation whether the resulting measurement precision of the gain scores was small enough to make the scores meaningful. In any case, given that gain scores are inherently much less precise than the component scores, one must start with very precise measures to obtain a precise gain score.

Whether adaptive testing can yield useful and informative gain scores can only be answered on a case-by-case basis. Item pool quality and overall test length will vary in each potential setting and, consequently, so will test precision. Kang and Weiss (2007) found that an item pool of highly discriminating items with a difficulty span greater than the range of true theta worked best in measuring individual change. For example, the 35-item test length of the simulated CAT in the current study may have been too short to provide the precision needed for accurate gain scores. A longer test, however, might require a broader and deeper item pool to be fully effective and thus be difficult for some assessment programs to support. Studies based on realistic settings (i.e., a CAT vertical scale and existing item pool), are needed to determine whether adaptive testing can provide precise gain scores for vertical scales.

Though CAT designs are attractive for a variety of reasons, there are also a few roadblocks that stand in the way of using adaptive testing for NCLB. In many school districts across the country the computer laboratory facilities are inadequate to support moving to a fully computerized statewide testing program. The situation is improving as time goes on, however, and many states currently have computer-based statewide tests of some form or another. In those states where computer test is currently a viable option, there remain the challenges of migrating from computer-based to fully adaptive tests, namely logistic, psychometric, and cost issues. Critical concerns such as test security, item bank development and maintenance, and score comparability, among others, have been fully discussed in the CAT literature. The feasibility, as well as the advisability, of a using CAT for accountability purposes must be addressed on a case-by-case basis. Nonetheless adaptive testing has been successful in some settings.

Beyond the issues of hardware availability and logistical issues, a key reason CAT is not widely used for statewide accountability tests is that NCLB rules currently mandate that test questions be on-level for each grade. This restriction negates a potential advantage of using adaptive testing, namely that a CAT item pool would be function best by spanning across grade levels. A single pool spanning grades would not only potentially allow for better measurement of low and high proficiency students, but it would also likely improve the stability of the vertical scale linking the grade-level scales together.

One state that has implemented an adaptive model for accountability is Oregon, which administers computer adaptive tests in several subjects (Oregon Department of Education, 2008). Due to the NCLB restriction stated above, however, Oregon item pools are strictly on-level for the grade being tested. Others have put forth arguments for using adaptive testing for accountability purposes (e.g., Kingsbury & Hauser, 2004; Steering Committee of the Delaware Statewide Academic Growth Assessment Pilot, 2007). Although the current legislation is not favorable to adaptive testing, the potential advantages offered by adaptive testing warrant exploration. The computer simulation described in this paper is such an exploration.

Further study of how to best create IRT vertical scales is another important area for research. Gain scores are only meaningful to the degree that the model forming the underlying scale is accurate. How to best create an appropriate vertical scale, and whether this can be done with a unidimensional scale, remain important research questions. Certainly, there are researchers who are pessimistic about the usefulness of vertical scales (e.g., Schafer, 2006). A major conclusion from the current study, however, is that even if the psychometrics challenges of constructing valid IRT vertical scales are overcome, gain scores from such scales may be too imprecise to be meaningful. While a poorly constructed vertical scale clearly cannot be expected to yield useful scores, a well-defined vertical scale in and of itself does not guarantee that reported individual scores will be precise enough to be support meaningful decision-making.

References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord & M.R. Novick, *Statistical Theories of Mental Test Scores* (pp. 395–479). Reading, MA: Addison-Wesley.

Fischer, G. H. (2003). The precision of gain scores under an item response theory perspective: A comparison of asymptotic and exact conditional inference about change. *Applied Psychological Measurement. 27*, 3-26.

Ho, A. (2007). Growth models under NCLB: Back to basics. *NCME Newsletter, 15*(4), 5-7.

Kang, G. K., & Weiss, D. J. (2007, June). *Comparison of Computerized Adaptive Testing and Classical Methods for Measuring Individual Change*. Paper presented at the GMAC Conference on Computerized Adaptive Testing, Minneapolis, MN.

Kingsbury, G.G. & Hauser, C. (2004, April). *Computerized Adaptive Testing and No Child Left Behind*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

May, K., & Jackson, T.S. (2005). IRT item parameters and the reliability and validity of pretest, posttest, and gain scores. *International Journal of Test*ing, *5*, 63-73.

May, K., & Nicewander, W. A. (1998). Measuring change conventionally and adaptively. *Educational and Psychological Measurement, 58*, 882-897.

Mellenbergh, G. J. (1999). A note on simple gain score precision. *Applied Psychological Measurement*, *23*, 87-89.

Nicewander, W. A. (1991, May). *The Conditions Under Which Gains in Achievement can be Accurately Measured and a Reliability-enhancing, Non-linear Transformation for the*

Ordinary Difference Score. Paper presented at the Model Based Measurement Workshop, Educational Testing Service, Princeton, NJ.

Oregon Department of Education (2008). *2007-2008 Technical Report: Oregon Statewide Assessment System*. Retrieved November 21, 2008, from http://www.ode.state.or.us/search/page/?=1305.

Ostini, R., & Nering, M. L. (2006). *Polytomous Item Response Theory Models*. Thousand Oaks, CA: Sage.

Schafer, W.D. (2006). Growth scales as an alternative to vertical scales. *Practical Assessment, Research & Evaluation, 11*(4). Available online: http://pareonline.net/pdf/v11n4.pdf.

Singer, J. D. & Willett, J.B. (2003). *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. New York: Oxford University Press.

Steering Committee of the Delaware Statewide Academic Growth Assessment Pilot (2007, October). *A More Accurate Growth Model: Using Multigrade Adaptive Assessments to Measure Student Growth*. Retrieved November 21, 2007, from http://www.nwea.org/assets/weblinked/DLReport%202007_11.pdf.

Sympson, J. B., & Hetter, R. D. (1985). *Controlling item-exposure rates in computerized adaptive testing.* Proceedings of the 27th annual meeting of the Military Testing Association (pp. 973-977). San Diego, CA: Navy Personnel Research and Development Center.

Thompson, T. (2007, April). *Some Issues in Computing Conditional Standard Errors of Measurement for State Testing Programs*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.

Thompson, T., & Way, D. (2007, June). *Investigating CAT Designs to Achieve Comparability with a Paper Test*. Paper presented at the GMAC Conference on Computerized Adaptive Testing, Minneapolis, MN.

Van der Linden, W. J., & Glas, C. A. W. (Eds.). (2000). *Computerized Adaptive testing: Theory and Practice*. Boston, MA: Kluwer.

Wainer, H. (Ed.). (2000). *Computerized adaptive testing: A Primer (2nd Edition)*. Mahwah, NJ: Lawrence Erlbaum Associates.

Wang, W.-C., & Wu. C.-I. (2004). Gain score in item response theory as an effect size measure. *Educational and Psychological Measurement, 64*, 758-780.

Willett, J.B. (1997). Measuring change: What individual growth modeling buys you. In E. Amsel and K. A. Renninger (Eds.), *Change and Development: Issues of Theory, Method, and Application* (pp. 213-243). Mahwah, NJ: Lawrence Erlbaum Associates.

Williams, R. H., & Zimmerman, D. W. (1996). Are simple gain scores obsolete. *Applied Psychological Measurement, 20*, 59-69.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1999). *BILOG-MG: Multiple Group IRT Analysis and Test Maintenance for Binary Items* [Computer program]. Chicago, IL: Scientific Software International.