

Growth, Precision, and CAT: An Examination of Gain Score Conditional SEM¹

Tony D. Thompson²

Pearson

Background

Measurement of student growth 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 measurement. As indicated by Ho, interest in growth modeling is surging and is likely to continue to do so, especially given draft reauthorization proposals for the No Child Left Behind Act (NCLB) that prominently feature growth models.

In the midst of the growth score surge, however, it also has been recognized that tests built for school accountability under the NCLB status model are not likely to provide ideal measures under a growth model (e.g., Steering Committee of the Delaware Statewide Academic Growth Assessment Pilot, 2007). Most nonadaptive proficiency tests, such as those designed to assess 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 with non-central proficiency will be less precise than those in the center. Despite the recent phenomenal interest in growth modeling, the precision (or lack thereof) of growth scores seems to have attracted less attention. It seems fruitful to examine growth score precision in order to provide reasonable limits on what can be expected from nonadaptive tests. Furthermore, it is useful to explore whether adaptive tests might substantially increase the precision of growth scores.

In this paper, the precision of the most basic of growth measures, the simple gain score, is examined within an IRT framework. It is shown that the conditional standard of error of measurement (CSEM) of the gain score is calculated as a simple function of the CSEM of the two component scores from which the gain score is derived. Further, an example of a vertical scale is developed using item parameter estimates from a state testing program. The example demonstrates the variation in measurement precision that can be anticipated when reporting gain

¹ Paper presented at the 2008 annual meeting of the National Council on Measurement in Education, New York, NY

² email: tony.thompson@pearson.com

scores. 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 nonadaptive 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.

Gain Scores

Scores resulting from taking the difference in 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 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. The difference between measures is taken as the effect of the treatment.

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 the student's score from the previous year to the current year's score for the student. Once a common scale has been 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 construct 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 far from ideal measures 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. 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.

A long-standing controversy in the psychometric literature has brewed over whether gain scores are inherently unreliable. Most of the literature examining this issue has focused on classical test theory representations. That is,

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

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

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

where S, T, and E designate the observed, true, or error score respectively, G is the gain score, and the subscript refers to either the first or second testing occasion. Equation 3 shows that both true and error components are subtracted. Subtracting true scores generally diminishes true score variance, but subtracting uncorrelated error scores adds to error score variance. This combination can virtually eliminate reliability (but see Williams and Zimmerman, 1996, who show this result is not inevitable).

The current paper takes the point of view of that espoused by Mellenbergh (1999) and Fischer (2003), namely that while reliability can be a useful, it is a relative rather than an absolute indicator of a test's measurement precision. By definition, reliability is the ratio of true score variance to total variance, and can be viewed 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, but the reliability of the scores for the two groups could be markedly different due to true score variance differences between the groups.

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

The IRT conditional standard error of measurement (CSEM) is an absolute measure of test precision for a given score scale. It has the further advantage of varying by true proficiency, rather than just being a single value that summarizes the overall test. For tests that follow an IRT model, overall classical test reliability becomes a less important indicator. For gain scores derived from an IRT model, such as scores derived from an IRT vertical scaling model, it makes sense to examine measurement precision conditionally on true proficiency rather than to rely on a single population dependent number to characterize growth score precision.

Despite the attractiveness 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 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 3PL model (Birnbaum, 1968) and its polytomous generalizations. 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 inducing a ceiling effect. They showed that IRT scoring possibly combined with adaptive testing could reduce scale distortion for gain scores. One important implication of this study for the present context is that it is crucial that the vertical scale is well formed and appropriate for the application. A poor vertical scale will likely produce poor gain scores.

In another paper, 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 and 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 very small item reliabilities ($< .05$ 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 paper did, the current study argues that using the IRT CSEM is a more effective and powerful tool in studying gain score precision.

Gain Score CSEM

In this section of the paper, the gain score conditional standard error of measurement is derived. Following the derivation is a demonstration of what the gain score CSEM might look like for an IRT vertical scale. The context for this demonstration is an IRT vertical scale for a state testing program that links adjacent grades from 3-8.

When IRT is the model underlying the vertical scale upon which gain scores are based, the CSEM of gain scores follows the same definition as the CSEM for any score. That is, the 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.

Take as an example the case of gain score G , defined as the difference between two scores from two occasions but scaled to a common metric. Hence,

$$G = S_2 - S_1, \quad (4)$$

where the subscript refers to either the first or second occasion. For a given pair of theta values θ_1 and θ_2 , assume that the error from occasion one is uncorrelated with the error from occasion two. The conditional error variance of the gain score is then given by,

$$Var(G | \theta_1, \theta_2) = Var(S_1 | \theta_1) + Var(S_2 | \theta_2) \quad (5)$$

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,

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

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 natural to condition on the true gain. That is, let

$$\eta_i = \theta_{2i} - \theta_{1i}, \quad (7)$$

where η_i is the true gain for student i . A particular student's gain score averaged across replications will equal their true gain. However, the standard deviation across replications for

the student will not necessarily equal the replication standard deviation for other students who have the same true gain. Students with the same true gain may have proficiencies at different points of the theta distributions. Thus, conditioning must occur at the student level, rather than conditioning on 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 are duplicated here. The two reading comprehension tests used were from grades 3 and 4. 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. The actual vertical scales developed for the state in question is not used here. Instead, a slightly simplified version is described 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. The CSEM for each grade was computed using methods given in Thompson (2007). Once the 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 slightly 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. The graph shows that as theta values went into the positive range for both grades, the CSEM increased as well. The maximum CSEM values were found when both grade 3 and grade 4 thetas were greater than 1.4. The CSEM values for this region were approximately 1.4.

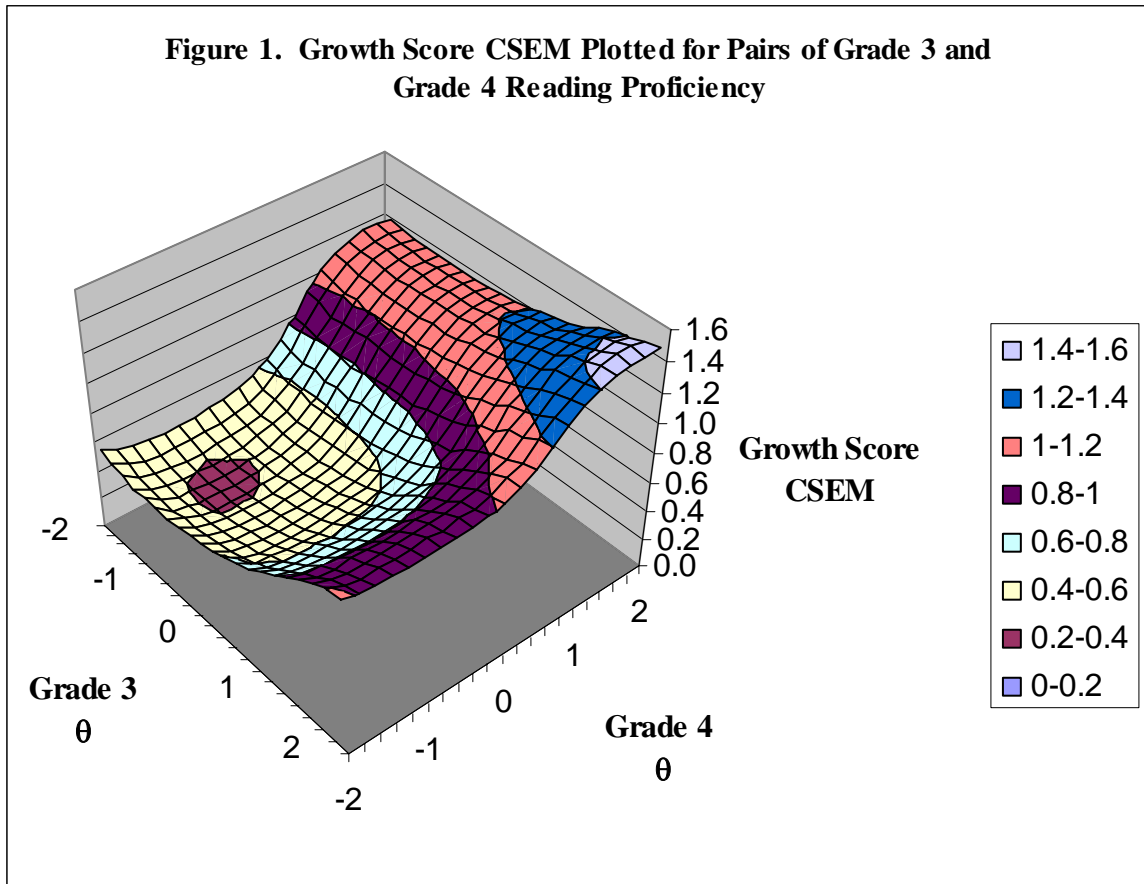
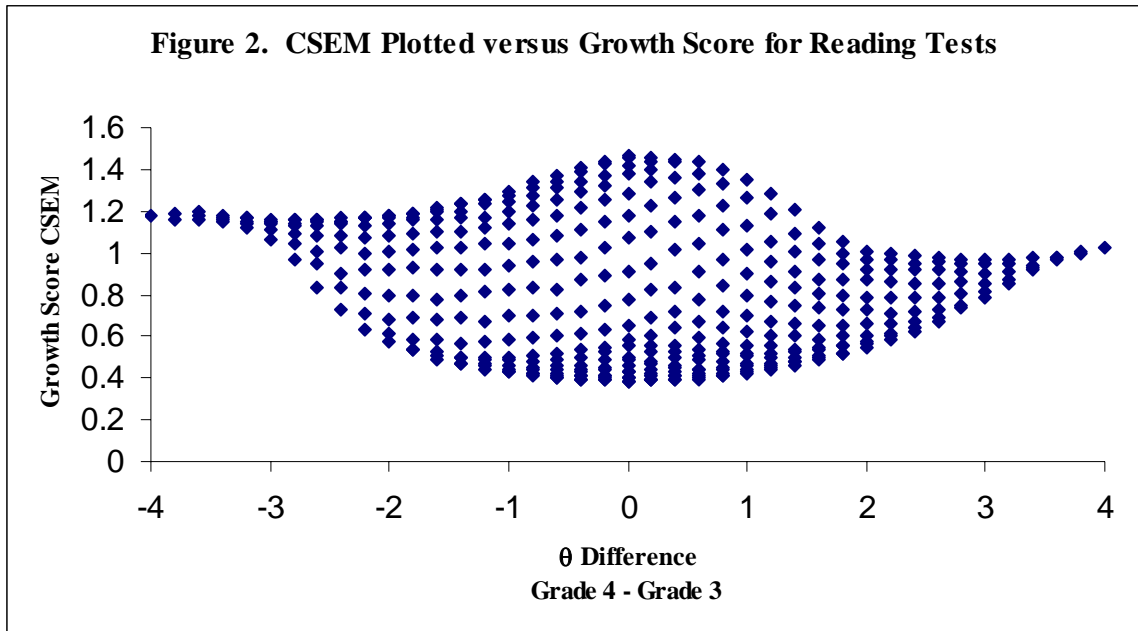


Figure 1 can be explained by the fact that the tests for both grades had the best measurement precision in the region of the theta scale from approximately -1.2 to -.6. Both were relatively easy tests for their respective populations. Because the on-grade tests worked best for this region of the theta 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.

Another way to look at the gain score CSEM is given 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 in the plots, the CSEM is not constant for a given true gain score, because the same true gain score can be obtained from many 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.



The other striking aspect of the graphs is that the minimum CSEM is around .4. This means that 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 only created for this paper 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 interesting comparisons between individual may 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 indeed grow.

The overall conclusion from Thompson (2007) was that a vertical scale developed from a typical K-12 testing program might be able to report informative gain scores for some students,

but that for a large proportion of students gain scores would be non-informative. The finding stems directly from the reality that individual grade-level tests are not designed to measure each student with equal precision. Equating 6 shows that the gain score measurement precision can be no better than the precision of either of the two component that contribute to the score, and thus, a component score of low precision necessarily results in gain score of low precision. In looking at ways to address this issue, a natural way to obtain high precision of measurement across the distribution of proficiency is through adaptive testing. 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 be able to obtain reliable growth scores 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 CAT or conventional test at two points in time. They defined significant change being observed whenever a measurement procedure yielded non-overlapping error bands for the two testing occasions. They found that the conventional test measured change best for the proficiency levels that the test was targeted to. The CAT, however, measured change equally well across the proficiency scale. In addition, the CAT was superior to the conventional approaches in measuring change.

The remainder of the current paper is devoted to describing a small-scale study that investigates the potential of CAT in the context of a vertical scale to obtain precise gain scores across the proficiency distribution. The study also further highlights how the CSEM function from Equation 6 can be used to evaluate the precision of IRT-based gain scores.

CAT Gain Score Precision

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. For those states where computer test is currently a viable option, though, migrating from computer-based to fully adaptive tests adds further 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 using CAT for accountability purposes testing will have to be addressed on a case-by-case basis. Suffice it to say that experience has shown adaptive testing to be successful in some, but not all settings.

Beyond the issues of hardware availability and logistical issues, though, a key reason CAT is not 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. Others have made the argument 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 below is such an exploration.

CAT Simulation Method

The computer simulation compares the gain score precision from paper and adaptive versions of a mathematics test. Because there are currently no NCLB accountability adaptive tests for reasons cited above, a detailed simulation of a vertical scale of such a 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. In that study, the authors compared several different CAT designs and compared resulting score comparability with a paper version. Although the purpose of that study was different and no vertical scale was simulated, it was deemed valuable to use a realistic CAT simulation rather than use mocked-up item parameters and hypothetical content constraints. The Thompson and Way simulation provided a realistic CSEM function for the CAT and paper test comparison. Here, we arbitrarily call this the “upper” grade CSEM. For the lower grade CSEM, a few assumptions were made. It was assumed, like for the grades 3 and 4 reading tests described previously, that the tests from two adjacent grade levels would be similarly targeted from their respective populations. CAT item pools for the two grades were likewise assumed to be of similar breadth and depth for their respective populations. Further, 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 as that 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). Essentially, 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.

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 but 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

Three parameter logistic (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.

Content Balancing Method

Content was balanced for the 10 objective score areas described in Table 1. The goal was for each objective to be proportionally represented in the CAT the same as the paper test. The algorithm picked the “most needy” content area at each point in the CAT (ties resolved randomly). Most needy meant the objective whose proportional representation was most dissimilar to the paper test content distribution. The items from the content area defined as “most needy” were the only items eligible for use by the item selection method.

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, and theta moves 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 Sympon-Hetter exposure control procedure was implemented (Sympon & 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.

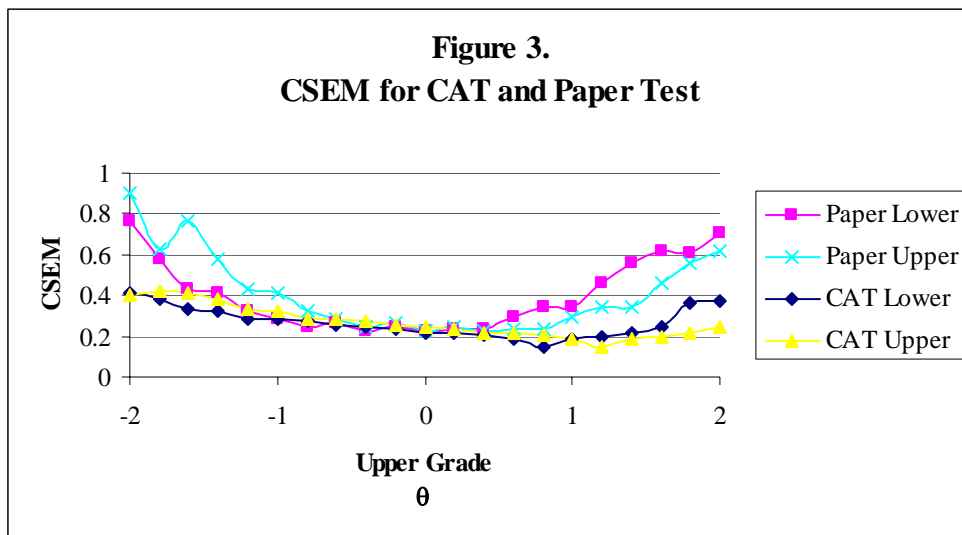
Other Simulation Details

Simulated response vectors using 41 true proficiency values from -4 to +4 were randomly generated. At each proficiency level, 200 simulated examinees were generated. After completing the initial simulation, two more replications were performed.

Results

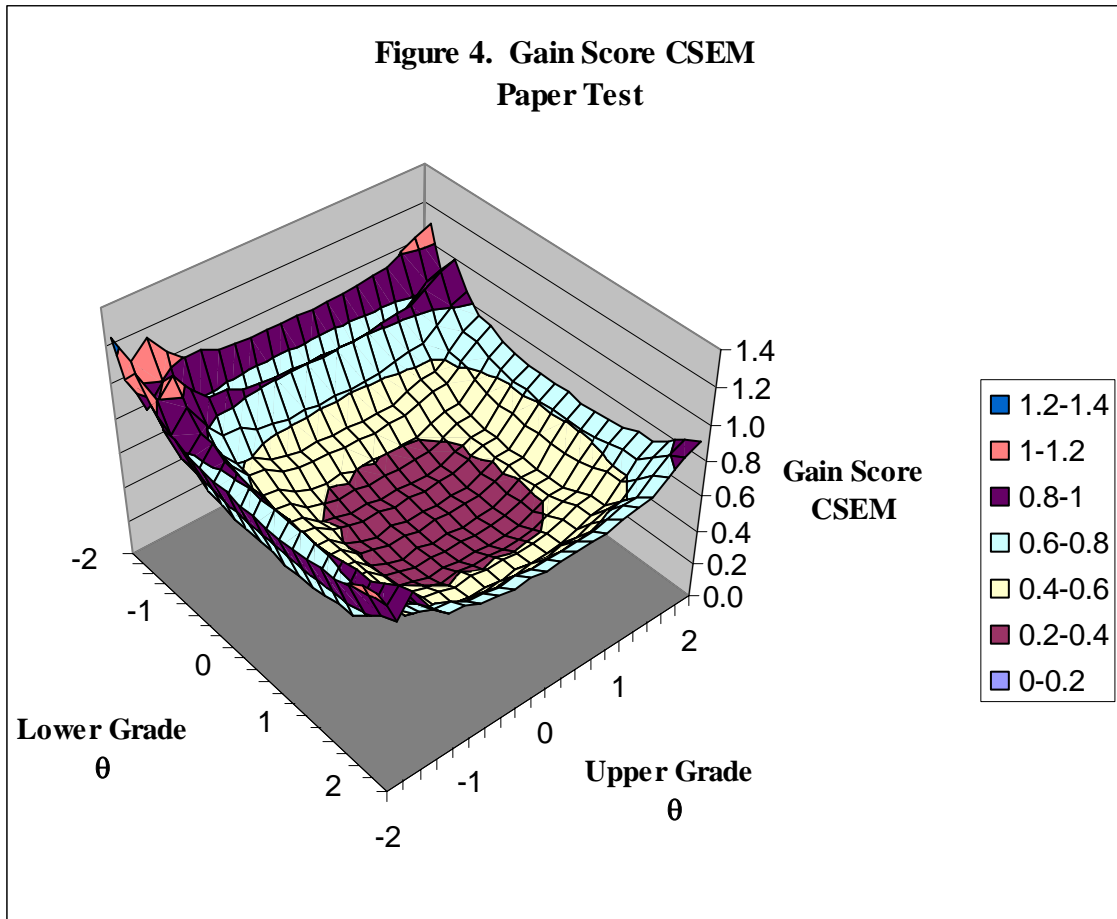
The same pattern of results was found in each of the three replications. For the purpose of simplifying the presentation of results, the tables and graphs below report the first of the three replications performed. Thompson and Way (2007) report on the measurement quality of the 35-item CAT test as compared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



The gain score CSEM functions are presented in Figure 4 (paper version) and Figure 5 (CAT version). The paper test version show some similarities to the vertical gain score from the reading test given in Figure 1. There is a wide range of CSEM values depending upon where the

student might start in the lower grade and where they end up in the upper grade. Unlike for 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 for the more extreme values of theta does the CSEM function become large.



In contrast, for the CAT (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.

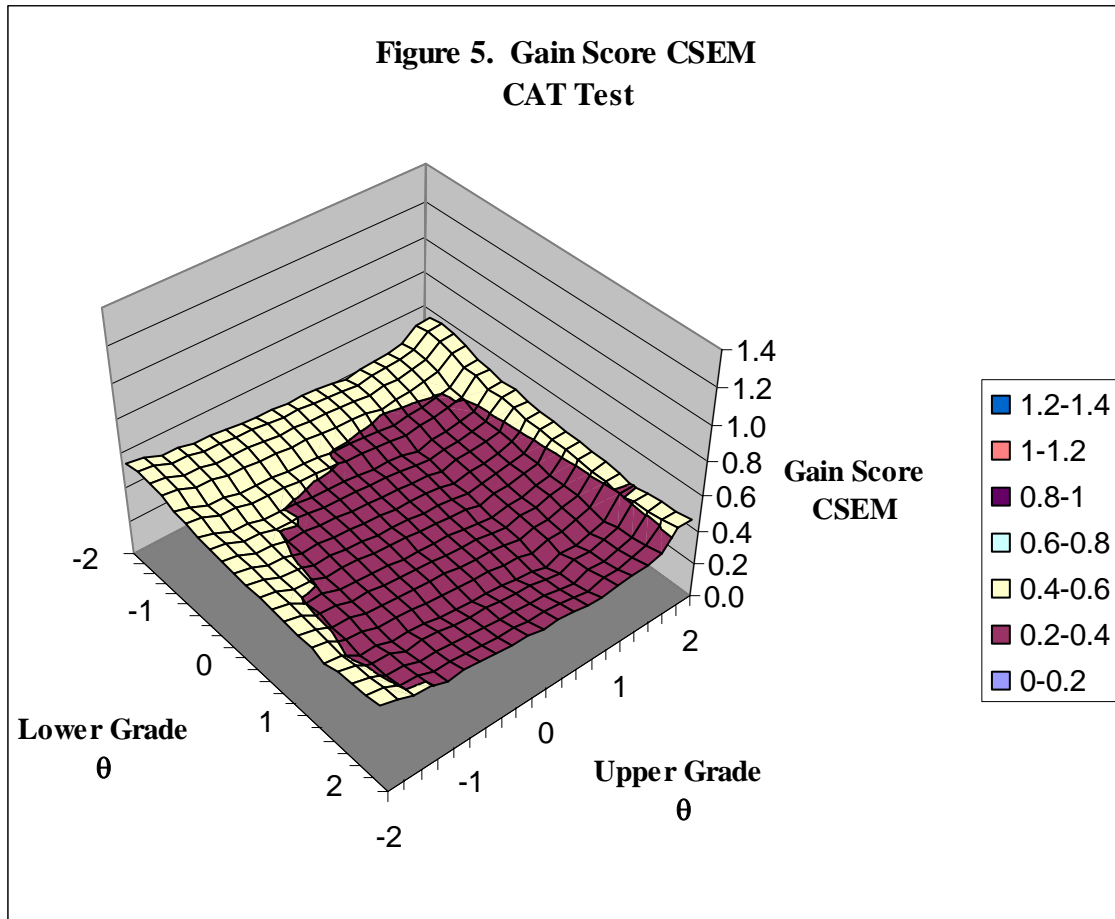
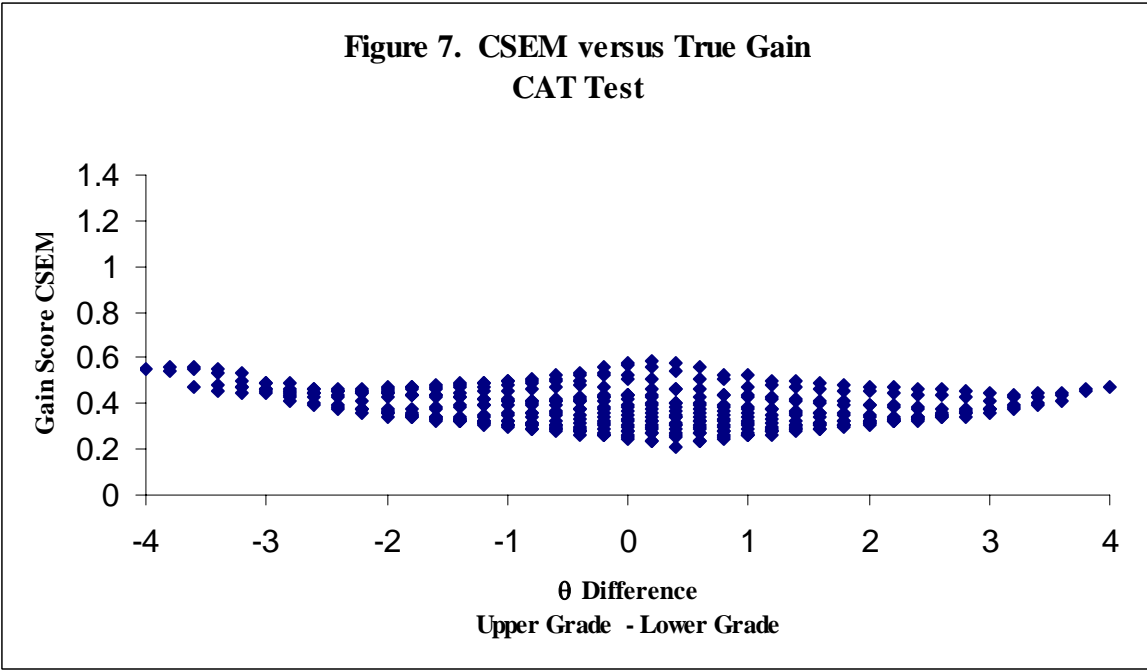
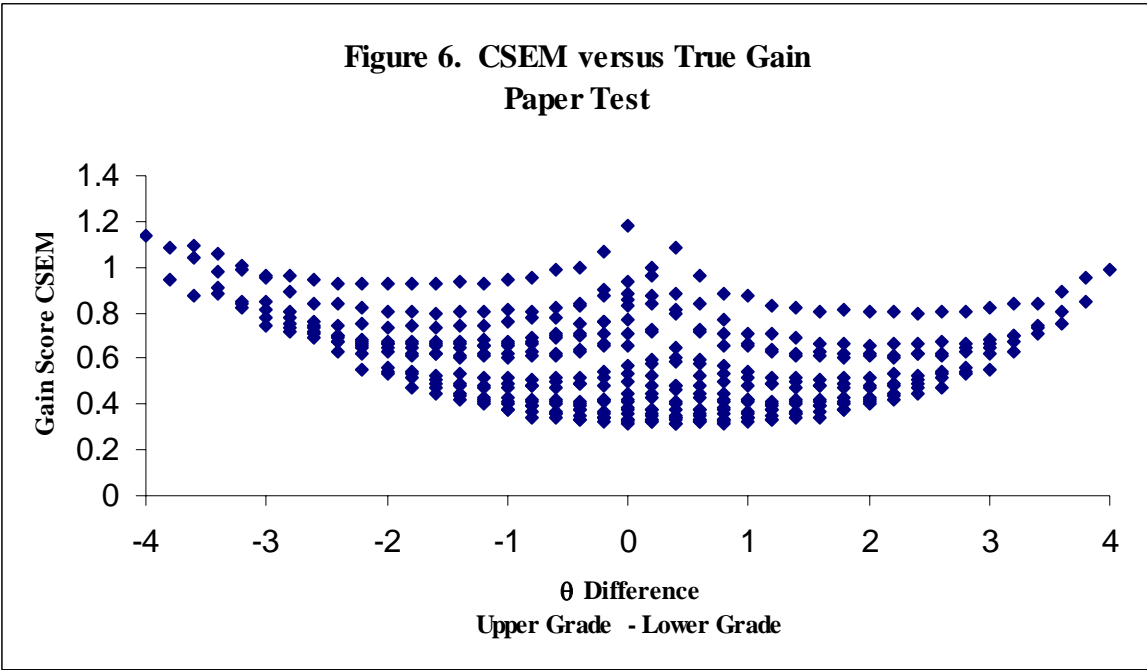
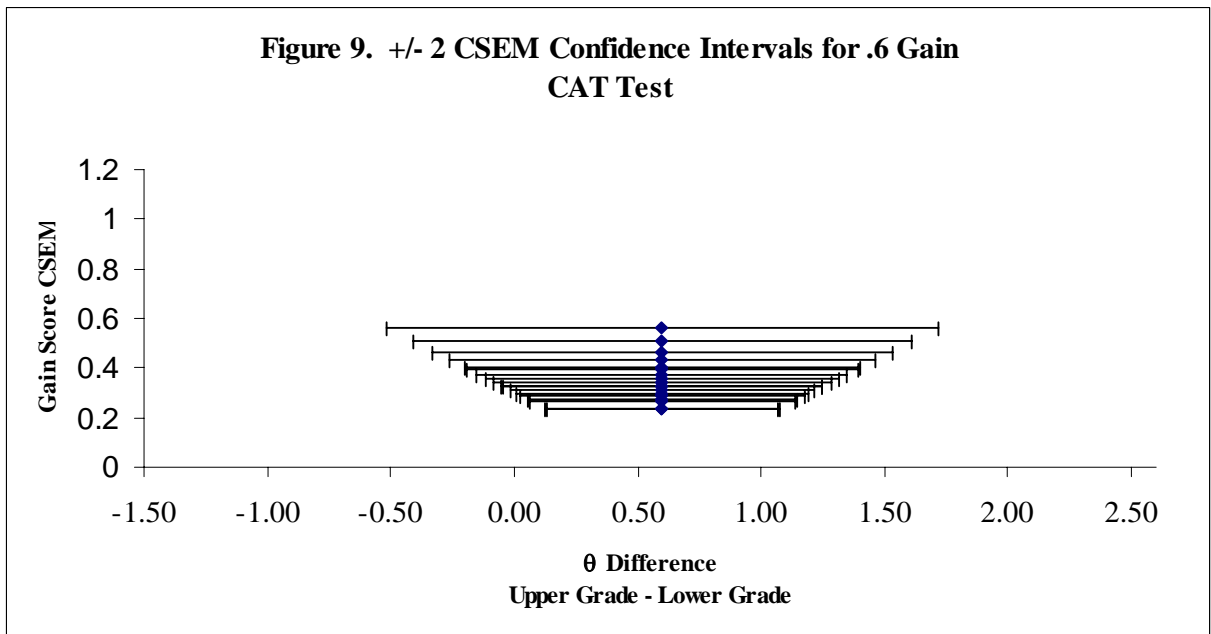
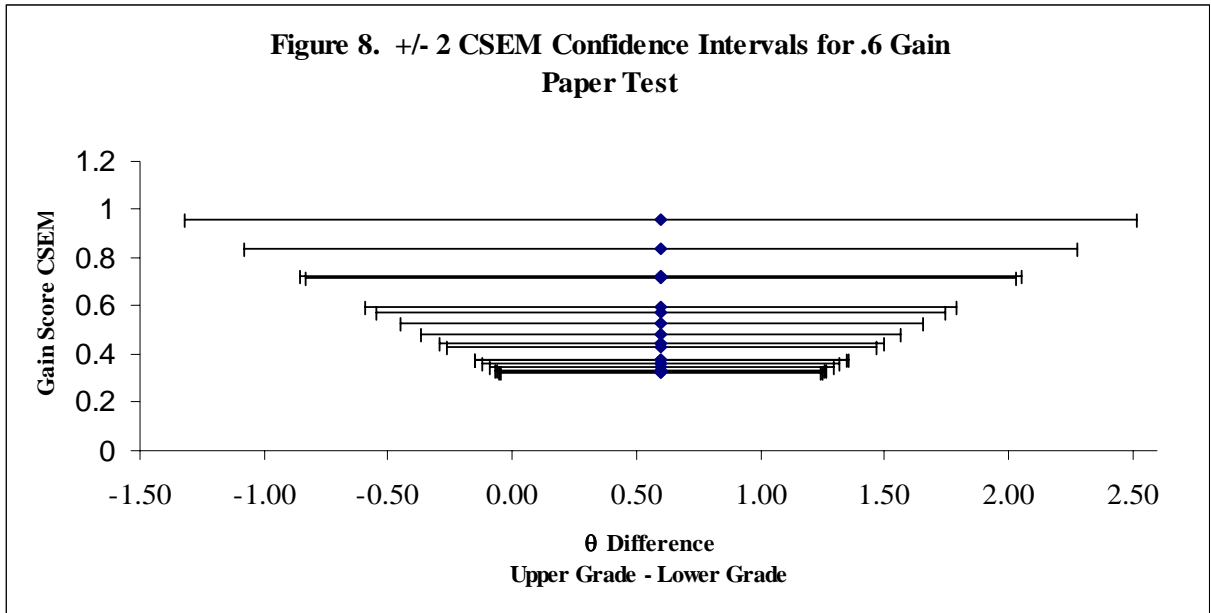


Figure 6 and Figure 7 present the gain score CSEM plotted against the true gain. In these plots, the lower end is not relevant, as we do not expect many students to have a substantial negative gain. It is perhaps in the 0 to 2 region that we expect the vast majority of student gain scores. Although the 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.



To investigate the variability in CSEM values more deeply, ± 2 confidence intervals were formed for the gain score of .6. 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.



As is consistent with the other results examined, the paper test gives similar 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 paper test intervals seem too wide to make the gain score meaningful. 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.

Conclusion

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 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 must be less precise than the lower and upper grade measures. Furthermore, CSEM values can vary for two individuals with the same true gain.

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 had the attention from research studies that the topic deserves. Before the final decision is made about what growth score to use for a testing program, it must first be determined if the reported scores will be precise enough to be meaningful. This point is underlined in the reported simulations by the often poor precision of the paper test gain scores. These results agree with much of the long history of research on the unreliability of difference scores. However, the previous research focused 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. A further point is that if gain scores are large relative to the error of measurement, then gains scores are meaningful even if the error is fairly large in an absolute sense. That is, if everyone gains 3 points on the theta scale and the error is only 1 theta point, then we can be fairly confident that students really did gain, even though the absolute error is fairly large. A study showing typical observed gains for an operational vertical and comparing the gains to the CSEM would be interesting. 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 often much smaller and that were also much more consistent in magnitude than the paper version of the test. However, it was not clear from the

simulation whether the resulting measurement precision of the gain scores was small enough to make the scores meaningful. 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. Developing such a pool could be difficult for some assessment programs. For the current study, it could be that the 35-item test length of the simulated CAT was too short to provide the precision needed for accurate gain scores. A longer test length, however, might require a broader and deeper item pool to be fully effective. Studies based on realistic settings (i.e., a CAT vertical scale and existing item pool), will be needed to determine whether adaptive testing can provide precise gain scores for vertical scales.

Further study of IRT vertical scales is another important area for research. Gain scores can only be 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., Shaffer, 2006).

Finally, although this paper dealt strictly with the CSEM, it may be that the asymptotic confidence intervals formed by using the CSEM may not be ideal. That is, the CSEM is useful for identifying which areas of the proficiency distribution where the test measures well (making an inference from a true score), but it is less useful when making inferences from observed scores. The use of confidence intervals based on prediction models is an area where there has been little research for K-12 tests. Further research on creating better confidence intervals might shed more light on the potential usefulness of gain scores.

References

- 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, December). Growth models under NCLB: Back to basics. *NCME Newsletter*, 15(4), 5-7.
- Kang, G.K., & David J. 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.
- 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.
- May, K., & Jackson, T.S. (2005). IRT item parameters and the reliability and validity of pretest, posttest, and gain scores. *International Journal of Testing*, 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.
- 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 from NWEA website:

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.
- 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*. Mahwah, NJ: Lawrence Erlbaum Associates, Chapter 11, 213-243.
- 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: Scientific Software International.