DOCUMENT RESUME

ED 453 266 TM 032 803

AUTHOR Sykes, Robert C.; Truskosky, Denise; White, Hillory
TITLE Determining the Representation of Constructed Response Items in Mixed-Item Format Exams.
PUB DATE 2001-04-00
NOTE 42p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Seattle, WA, April 11-13, 2001).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Constructed Response; Elementary Education; *Elementary School Students; Error of Measurement; Item Response Theory; Mathematics Tests; *Reliability; Scores; Test Construction; Test Format; *Test Items; Writing Tests
IDENTIFIERS Unidimensionality (Tests); *Weighting (Statistical)

ABSTRACT

The purpose of this research was to study the effect of the three different ways of increasing the number of points contributed by constructed response (CR) items on the reliability of test scores from mixed-item-format tests. The assumption of unidimensionality that underlies the accuracy of item response theory model-based standard error predictions of reliability was initially evaluated for these tests. Large samples of students who had taken mixed-format field tests in mathematics at grades 5 and 8 and writing at grades 3 and 8 were available from a state criterion-referenced testing program. The selection of subsets of items from test-blueprint-representative forms of similar content and difficulty permitted an evaluation of the effects of weighting CR items on total test scores relative to criterion scores of putatively greater generalizability. As expected, there was a cost in terms of precision of having fewer, though weighted, CR items across a wide range of ability. The increment in standard error attributed to weighting was predictably less in the middle of the scale where the forms were targeted. The magnitude of the increase in error and the particular portion of the scale where it occurs are determined by the locations and amount of information contributed by the deleted CR items relative to those that are retained. Implications of different approaches to weighting are discussed. (Contains 5 tables, 10 figures, and 10 references.) (SLD)

Determining the Representation of Constructed Response Items in Mixed-Item Format Exams

Robert C. Sykes

Denise Truskosky

Hillory White

CTB/McGraw-Hill

This paper was presented at the Annual Meeting of the National Council on Measurement in Education in Seattle, April 2001.

INTRODUCTION

Constructed response (c.r.) items are now frequently found complementing multiple choice (m.c.) items in mixed-format examinations. These items are believed important in their capability to influence curriculum through their assessment of skills not evaluated by m.c. items, such as organized or creative expression, while the m.c. items allow a breadth of content coverage by an evaluation of content or factual knowledge. The employment of IRT models allows both types of items to be scaled together, providing the advantages of a single score if the assumptions of the model such as unidimensionality are met. Traub (1993), in a review of the studies that existed at that time, suggested that the items of the two formats probably do not measure different characteristics for tests in the Quantitative or Reading Comprehension domains but may measure different characteristics for Writing.

The use of both the c.r. and m.c. item formats requires a determination of the degree to which they will be represented or weighted. One manner of defining the contribution the c.r. items will make to the total test score, as well as that of the m.c. items, is through the items' psychometric characteristics. Specifically, the use of IRT (pattern) scoring implies that a decision has been made to weight each item by its reliability (i.e. discrimination). This type of psychometrically imposed weighting, resulting in total test scores that are optimal in terms of reliability, may be contrasted to the test-designer

imposed weighting of item formats that is the subject of this research. Because a set of c.r. items is not likely to produce a total score with reliability as great as a set of m.c. items administered in the same period of time (Wainer & Thissen, 1993), a rationale for test-designer imposed weighting would presumably be that they are desired to increase the validity of the examination.

Three different types of test-designer imposed weighting utilizing number-correct scoring with the employed IRT model are possible. (The assignment of the worth or point value of each type of item is another method of weighting items that is not considered here.) The first of these methods of weighting is through the specification of the test blueprint (i.e. blueprint representation). The representation of c.r. items in a test (i.e. relative proportion of total score points contributed by the c.r. items) is determined through this method by the stipulation of the number of c.r. items required in those categories assessing skills that can only be evaluated by these items and the number of c.r. items from categories that can be evaluated using either c.r. *or* m.c. items.

The number of c.r. items in these latter categories can vary depending upon the availability or desirability of c.r. items. Relatively large numbers of c.r. items may be necessary for a test if there are many categories of the former type and/or c.r. items are preferred to fill the latter type of blueprint categories.

Because c.r. items generally require longer response times, however, it may not be feasible to administer as many as are desired within the time available for testing. Testing time is especially a problem when the c.r. items require an extended response (e.r.), such as the writing samples given in response to a prompt. It may not be possible to administer more than one of these e.r. items, along with the accompanying m.c. and other c.r. items.

Although administering a larger number of e.r. or c.r. items would be desirable from the standpoint of the generalizability of test scores, it is possible to increase the number of points coming from a set of c.r. items without increasing their number (and testing time). A second possible type of weighting is implemented by multiplying the portion of the test characteristic curve (tcc) that is contributed by these items by an integer factor (i.e. tcc component weighting). Thus if it was desired to increase the number of points contributed to the total test score by a single e.r. response from six to 12 points the expected e.r. score would be multiplied by two. The increased expected item score is then added to those for the other items to obtain the expected total raw score for scale scores across the scale and thus the scoring tables.

Ito and Sykes (2000) examined the effect of weighting sets of c.r. items through the test characteristic curve relative to a criterion of no weighting for three Writing tests. The authors documented relatively small decreases in the precision of test

scores when a limited number of c.r. items were weighted.

A third way of increasing the representation of c.r. items is the summing, rather than averaging (and if necessary rounding to the nearest integer), of the ratings of two readers (i.e. summed readings or ratings). In addition to the point value of the item being doubled the number of score levels for each c.r. item is increased from n (the number of levels of the rubric including 0) to 2n-1. Summed ratings is more restricted than tcc component weighting in that it requires multiple readers for each c.r. response and hence is limited to increasing the points from the c.r. items by a factor of two without prohibitively increasing the number of raters(and readings).

The method of summed ratings is imposed through the item parameter estimates and thus the latent scale. In contrast tcc component weighting is implemented through the score obtained after the set of c.r. items, with their rubric-determined point values and number of levels, has been scaled with the m.c. items. Because the number of levels of the c.r. items is increased with summed ratings item reliability may change, potentially affecting form reliability and IRT test score information.

The purpose of this research was to investigate the effect of the three different ways of increasing the number of points contributed by the c.r. items on the reliability of test scores from mixed-item-format tests. The assumption of unidimensionality that underlies the accuracy of IRT model-based standard error predictions of reliability was initially evaluated

for these tests.

METHOD

Source Data

Large samples of students that had taken mixed-format field tests for Math at Grades 5 and 8 and Writing at Grades 3 and 8 were available for a state criterion-referenced testing program. Responses to the subset of items in each of the field test forms that were later chosen to constitute a complete operational form were selected. Consequently the selected items for each grade/content area (hereafter forms) represent the operational test blueprints.

Responses to a second prompt were included with each of the two Writing forms. Although an item score for an extended response to a prompt is computed as an average over a number of analytic traits in the testing program, the score on a single trait - Organization - was utilized in these analyses.

Only students who responded to at least 2/3's of the selected items were used. Omits were treated as not correct.

The number of scored items and their point values (maximum number of points) are summarized below.

Content Area	Grade	Multiple Choice	Constructed Response: Two Point	Constructed Response: Six Point	Total Items	Total Points
Math	5	35	10	0	45	55
Math	8	35	10	0	45	55
Writing	3	29	3	2	34	47
Writing	8	25	6	2	33	49

Analyses

Construction of Forms

The subsets of items chosen for the operational tests represented a (unweighted) *Baseline* condition of test-blueprint representative forms, assuming that the addition of a second prompt to the two Writing tests would be required by the blueprint if testing time permitted.

Several different types of forms that weighted c.r. responses were created, each constructed to have the same number of total test points and approximate difficulty after weighting as the baseline forms from which the item responses were drawn. This was accomplished by partitioning c.r. items in a form into two matched sets of approximately the same difficulty (when the content and the number of the c.r. items permitted), deleting one of the sets, and weighting the remaining set.

Two instances of tcc component weighting were implemented. The first weighted the members of one of the sets of c.r. items in a form by a factor of two and is referred to as *CRx2*. The even number of c.r. items in the two Math *Baseline* forms (10) and the Grade 8 Writing form resulted in the matched sets being of equal size as well as similar content, with most frequently a content category of a deleted c.r. item being represented by a c.r. item in the remaining weighted set.

The second instance of tcc component weighting was based on the weighting of one of the two e.r. items in each of the two Writing forms by a factor of two and is referred to as *ERx2*.

The last type of weighting of the c.r. items, *Summed* Ratings, was created for those tests having c.r. items with more than two points (three levels including 0); that is, the two Writing tests. Only c.r. items with three or more points were subjected to a second reading and hence only the two writing prompts could have an item score based on a summed rating. One of the two prompts in each Writing form was deleted and a summed rating item score was obtained for the remaining prompt. Because the testing program called for a third, reconciliation reading if the two readers differed by more than a point, the item score was either a sum of two readings or the sum of three that was multiplied by 2/3's and rounded to the nearest integer.

Table 1 contains the items and their p-values (average item score divided by the maximum number of points) in the matched sets of c.r. items used in the creation of the *CRx2*, *ERx2*, and *Summed* forms of weighted c.r. responses.

Evaluations of Forms

Properties of the total test scores derived from the three types of forms, employing either tcc component or *Summed* rating weighting, were compared against the criterion baseline forms. The relationships between total raw scores and ability were examined through comparisons of tccs. Conditional standard errors were evaluated through standard error (se) curves. Scale scores produced by weighting were compared to those from the baseline forms and the magnitude of differences determined.

The dimensionality of the baseline forms was evaluated by utilizing Poly-Dimtest (Li & Stout, 1995) to detect violations of the assumption of unidimensionality. Specifically the presence of a significant dimension underlying the c.r. items was assessed.

Rating Process

Readers were trained to implement scoring rubrics; anchor papers, check sets, and read behinds were employed to verify and maintain scoring accuracy. Inter-rater reliability studies that incorporated second reads for a large sample of students taking each test indicated that the percentage of exact agreement on the c.r. items in the Math tests ranged between 92.58% and 100.00%. Exact agreement rates for the two-point Writing c.r. items ranged between 55.67% (66.46% for the second lowest exact rate) and 87.77%. The exact agreement rates for the selected "Organization" trait on the Writing prompts ranged between 58.84% and 62.23% with the approximate agreement rates (within one point) between 97.97% and 98.99%.

Scaling Process

Multiple-choice and open-ended items were scaled together using the generalized IRT model. With the generalized model a three-parameter logistic model (Lord, 1980) was used for the multiple-choice items:

$$P_i = P(X_i = 1|\theta) = c_i + \frac{1 - c_i}{1 + \exp[-1.7 A_i(\theta - B_i)]}, \qquad (1)$$

where A_i is the discrimination, B_i is the difficulty, and c_i is the lower asymptote or guessing parameter for item i.

A generalization of Master's (1982) Partial Credit model was used for the c.r. items. This two-parameter partial credit (2PPC) model is the same as Muraki's (1992) "generalized partial credit model." For a c.r. item with m_i score levels assigned integer scores that ranged from 0 to m_i - 1:

$$P_{ik}(\theta) = P(X_i = k-1|\theta) = \frac{\exp(y_{ik})}{\sum_{j=1}^{m_i}\exp(y_{ij})}, \qquad k = 1,\ldots,m_i \qquad (2)$$

where

$$y_{ik} = \alpha_i(k-1)\theta - \sum_{j=0}^{k-1}\gamma_{ij} ,$$

and $\gamma_{i0} = 0$. α_i is the item discrimination. γ_{ij} is related to the difficulty of the item levels: the trace lines for adjacent score levels intersect at γ_{ij}/α_i .

Parameter Estimation

Item parameter was conducted using the program PARDUX (Burket, 1991; 1995). Item parameters were estimated using marginal maximum likelihood procedures implemented with an EM algorithm. Evaluations of the accuracy of the program with simulated data (Fitzpatrick, 1990) have found it to be at least as accurate as MULTILOG (Thissen, 1986). The ability scale was defined by specifying a prior true θ distribution to have a mean of 0.0 and standard deviation of 1.0. Item parameter estimates were linearly transformed to a scale score metric by multiplying

by 50 and adding 500. The LOSS and HOSS (lowest and highest obtainable scale scores) were set for each form to allow for a wide range of scale scores that could accommodate different weightings of the c.r. items.

Student Scores

The relationship between the predicted raw score and the ability estimate θ_a (tcc) was obtained using the final item parameter estimates:

$$E(X_a \mid \hat{\theta}_a) = w_m \{ \sum_{i=1}^{mc} w_i P_i(\hat{\theta}_a) + \sum_{j=1}^{cr} w_j \sum_{k=1}^{m_j} (k-1) P_{jk}(\hat{\theta}_a) \} , \qquad (3)$$

where the predicted total score has been partitioned into components for the mc multiple choice items and the cr constructed response items. For (unweighted) number-correct scoring, such as that employed for the baseline forms, the weights w_i and w_j are all equal to 1.

Each selected c.r. item in the *CRx2* forms and selected e.r. item in the *ERx2* forms had w_j's set to 2, with again all w_i for the m.c. items set equal to 1. Scoring tables were constructed for all forms consisting of the scale scores corresponding to integer values of $E(X_a \mid \hat{\theta}_a)$.

The weight w_m, which multiplies each item probability along with the weights w_i or w_j, serves to determine the total number of points in the total score. Set to 1 the number of test score points is preserved at that for the baseline forms. If allowed to decrease between 1 and 0 the number of total score points can

be preserved even when c.r. items are weighted by factors (weights) that exceed two.

Information

The information of the raw score at ability θ is

$$I(\theta, \sum_{l} w_l X_l) = \frac{\left[w_m \sum_{l=1}^{n} w_l \sum_{k=1}^{m_i} (k-1) P'_{lk}(\theta) \right]^2}{\sum_{l=1}^{n} \sigma^2 (w_m w_l X_l | \theta)} . \quad (4)$$

The inverse of these values, plotted for the $\theta's$ across the ability continuum, constitute the standard error curves for the θ and corresponding scale score metrics.

Total information for each item was obtained by accumulating values of equation 4 over the range of ability.

RESULTS

Raw Score Statistics

Descriptive statistics for the *Baseline*, *CRx2*, *ERx2*, or *Summed* forms of the four tests are presented in Table 2. (Forms in the sense of differently scored versions of what may be the same set of test items.) The four *Baseline* forms differed in difficulty, with average p-values ranging between .375 for the difficult Math Grade 8 form and .686 for Writing Grade 3. Analyzing forms within meaningful comparison sets:

1) Math: *CRx2* versus *Baseline* for Grades 5 and 8 {Math (Two-Point) CR Analysis}

2) Writing: *CRx2* vs *Baseline* for Grade 8 {Writing CR (Two-Point) Analysis}, and

3) Writing: *ERx2* and *Summed* vs *Baseline* for Grades 3 and 8 {Writing ER Analysis}

reveals that the forms are very similar, an expected result given the relatively few items per forms that were weighted and the similarity in the difficulties of deleted and retained c.r. items.

The largest differences in form means within the three comparison sets was .33 for the *Baseline* and *Summed* forms for Writing Grade 8 (means of 28.63 minus 28.30, respectively). The largest difference from a *Baseline* standard deviation (sd) was .19 for the *ERx2* form for Writing Grade 8 (8.06 versus 7.87 {*Baseline*}, respectively).

The reliability (stratified alpha) of the *Baseline* form is consistently slightly above that of the *CRx2* forms, with the largest decrease occurring for Math Grade 5 (.871 versus .831). Test reliability is virtually the same across the *Baseline*, *ERx2*, and *Summed* Writing Grade 3 forms but is less for the *Baseline* Grade 8 Writing form (.868) than it is for *ERx2* (.894) and *Summed* (.892) versions. The relatively attenuated values for the stratified alphas for both Writing *Baseline* forms reflects the inability to include the retained (and weighted) prompt in the computation of the statistic for the *ERx2* and *Summed* forms. A strata size of only one item results in the e.r. item being excluded from the computation and subsequently higher stratified alphas for the weighted forms (i.e. forms with weighted c.r.

responses).

Dimensionality

To evaluate whether the c.r. items in the *Baseline* forms were dimensionally distinct from the m.c. items, Poly-Dimtest (Li & Stout, 1995) analyses were conducted using an AT1 subtest consisting of only c.r. items. The results of these analyses are shown in Table 3. All but one *Baseline* form, Math Grade 5, was found to be unidimensional. The Grade 5 Math *Baseline* form was marginally significant at p=.038.

Although the p-values for the c.r. items were generally lower than the m.c. items in each Math form, the AT1 subtests for both Math forms passed the Wilcoxon rank sum test as implemented in Poly-Dimtest using the default significance level of .02.

TCCs

Plots of the tcc's are presented, along with a tabling of the pairs of scale scores (SS) and predicted raw score (RS) values, for Math Grade 5 in Figure 1. Results for Math Grade 8 were similar and are not provided. Predicted scores for the *Baseline* and *CRx2* forms are very similar across the ability scale, differing by at most 1.39 raw score points (46.80 for *Baseline* versus 45.41) at a scale score of 625. The tcc's for the Writing Grade 8 CR Analysis in Figure 2 demonstrate even smaller differences between predicted scores with a maximum difference of .65 (43.54 for *Baseline* versus 42.89 for *CRx2*) at a scale score of 675.

The results for the ER Analysis for Writing Grade 8 presented in Figure 3 was similar to that seen for the *Baseline*, *Summed*, and *ERx2* forms for Writing Grade 5 (not presented). Predicted raw scores between the LOSS and HOSS for the *Summed* form differ by no more than 1.64 from the *Baseline* form (24.01 versus 25.65, respectively at 475) with even smaller differences between the *ERx2* and *Baseline* forms (max. difference of 13.34 - 13.08 = .26 at 400).

Standard Error

Total item information presented in Table 4 was preliminarily evaluated for the items in the four *Baseline* forms. The location of the items, that is the scale score value at which the item contributes the maximum information, is also provided. The mean information by item type at the bottom of the table indicates that the Math c.r. items contributes more than twice the amount of information, on average, than the m.c. items (e.g. .045 versus .021 for Grade 8).

The substantial information contribution of the Math c.r. items, exceeding the ratio of point values of the two item types (better than two-to-one), is not seen with the Writing c.r. items. The contribution of information by the Writing c.r. items is less than two-to-one for the two-point items and between approximately three-to-one and four-to-one (.068 versus .017 for Grade 8) for the six-point e.r. items. The information value for one of the e.r. items in the Grade 3 test (item # 33) is attenuated because the absence of students obtaining a perfect score of 6

necessitated a collapse of a category.

The *Baseline* se curves in the CR and ER Analyses depicted in Figures 4 through 7 are the plotted values of the reciprocal of item information (equation 4). In Figure 4 for Math Grade 8 (Math Grade 5 was similar and is not provided), the *CRx2* form demonstrates an 18% increase in standard error over baseline {(13 - 11)/11} in the 550 to 565 scale score range where precision is the greatest (hereafter point of form targeting). Scores for the *CRx2* form are slightly more precise (larger standard error) at the lower end of the scale but more than 30% less precise than the *Baseline* scores between 700 and 800 scale score points (e.g. {81-62(i}/62= 30.6% at 726 where the "i" indicates an interpolated value).

The CR Analysis of se curves for Writing Grade 8 in Figure 5 indicates error for the *CRx2* scores is larger than that for the Baseline form across the scale score scale, with the difference increasing after approximately 550. CRx2 scores have 21% greater error where the forms are targeted (23 versus 19 in the vicinity of 475). In the upper portion of the scale, the standard error for the *CRx2* scores has increased to more than 30% of that for *Baseline* (81 vs 62(i) at 726).

Figures 6 and 7 portray the ER Analyses for the two Writing forms. With the exception of intervals near the LOSS or HOSS of the forms *Summed* scale scores have a degree of error between that of scores for the *Baseline* and *ERx2* forms. At the point of targeting *Summed* and *ERx2* scores have standard errors at most two

scale score points (less than 11%) from that of the *Baseline* scores (21 for *ERx2* versus 19 for *Baseline* at 471 for Writing Grade 8 in Figure 7).

Error for the *ERx2* and *Summed* scores increase in the upper third of both scales. Relative to the Grade 3 *Baseline* se of 68 at 679 in Figure 6, the increased error is 44% (98{i}) and 19% (81{i}), respectively. At Grade 8 the increases, relative to a *Baseline* error of 61 at a scale score of 768, are 33% (81{i}) and 25% (76{i}), respectively.

Increased C.R. Item Weighting

By utilizing a value between 0 and 1 for w_m in equation 3 the relative weight applied to the c.r. items can be increased beyond a factor of two while preserving the same number of test points as the *Baseline* forms. The effect of increasing the relative weight of the retained e.r. item in the Writing Grade 8 test to a value of four times the weight of a m.c. item (*ERx4*) is depicted in Figure 8.

Standard error for *ERx4* scores is increased relative to the *Baseline* and other weighted forms. As is the case with the other weighted forms, the increment is relatively small in the lower portion of the scale (52 {i} vs 44 for an 18% increase at 349) but increases throughout the scale. Between 450 and 500, where the forms are targeted, the *ERx4* scores have 37% more error (26 vs 19) which increases to 47% at a scale score of 768 (90 {i} vs 61).

Scale Score Comparisons

Scale scores were obtained for the *Baseline* and weighted forms through unweighted and weighted raw score-to-scale score tables. Figure 9 contains plots (against *Baseline*) of the CR Analyses for the two Math tests and Writing Grade 8.

Scale scores obtained through weighting the retained c.r. items demonstrate a strong linear relationship to *Baseline* scores, with a product moment correlation (r) that exceeds .980 for both of the Math tests and a slightly lower .963 for Writing Grade 8.

Figure 10 depicts the relationship between the forms of the ER Analysis of the Writing Grade 3 forms, as well as scale scores obtained when weighting the retained e.r. item by a factor of four relative to a m.c. item (*ERx4*). Similar results, obtained for Writing Grade 8, are not presented.

Scores between the *Baseline* and the two weighted forms, *ERx2* and *Summed*, exhibit the high degree of correlation (.974 and .981, respectively) expected for forms that share all but one of their items, with no signs of non-linearity. *ERx4* scores have a slightly reduced correlation with *Baseline* scale scores (.942).

All the plots demonstrate greater scatter at the ends of the scale where error is greater. This is especially prominent at the upper portion of the Writing scales presented at the bottom of Figure 9 for Grade 8 and in Figure 10 for Grade 3.

Distributions of scale scores and their differences are described in Table 5, including those obtained after weighting the c.r. and e.r. items four times that of a m.c. item (*CRx4* and

ERx4). The means and standard deviations of the *CRx2* and *ERx2* scale score distributions resemble the corresponding raw score distributions in Table 2 in their similarity to the *Baseline* distributions.

Increasing the weight of the c.r. items by a factor as large as four (while maintaining the number of test points) serves to further increase the standard deviation of the scores relative to *Baseline* but generally not the means. This may be seen in the standard deviations for Writing Grade 8, which starting from a *Baseline* value of 58.15 increases with *CRx2* (63.75) and *CRx4* (70.09) as well as *ERx2* (60.69) and *ERx4* (65.25).

The similarity in the means of the weighted form distributions to *Baseline* reflect the comparability of the *Baseline* and reduced length forms containing the weighted c.r. items. Consequently the largest differences are between the *CRx2* and *CRx4* versus *Baseline* scale scores for the Grade 8 Writing forms (e.g. 502.60 for *CRx2* versus 499.69), which reflect the relatively larger difference in difficulty between the retained and deleted sets of c.r. items for this test (.524 vs .501, respectively, in Table 1).

Descriptive statistics for the differences between weighted form and *Baseline* scores are found in the right part of Table 5. Mean differences involving the *Summed*, *CRx2* and *ERx2* scores are small. The largest of these, 2.09 for *Crx2-Baseline* for Writing Grade 8, is inflated to a degree because of the difference in form difficulty mentioned above. Ten percent of the 3,288

students in this sample obtained a *CRx2* score that was at least 16 scale score points less than their *Baseline* scores (10%ile) while 10% received a *CRx2* scale score that was at least 21 points above their *Baseline* score. The next largest mean difference for *Summed*, *CRx2* or *ERx2* scores was a substantially smaller 1.03 for the *Summed* scores for Writing Grade 8. The 10^{th} and 90th percentile for this distribution of differences were -8 and 10, respectively.

An increase in the differences between weighted and *Baseline* scores as the weight given to the c.r. items increase can be seen when the *CRx4* and *ERx4* distribution of differences (relative to *Baseline*) is compared to the corresponding *CRx2* or *ERx2* distribution increase. For example, the *CRx4-Baseline* distribution of differences for Writing Grade 8 has a larger mean, sd, and more extreme 10^{th} and 90^{th} percentiles (5.29, 27.35, -27, and 37, respectively) than the *Crx2-Baseline* differences (2.09, 17.45, -16, and 21, respectively.

Discussion and Conclusions

The selection of subsets of items from test-blueprint-representative forms of similar content and difficulty permitted an evaluation, unconfounded by these factors, of the effects of weighting c.r. items on total test scores relative to criterion scores of putatively greater generalizability. As expected there was a cost in terms of precision of having fewer, though weighted (tcc component or *Summed*), c.r. items across a very wide range of

ability.

The increment in standard error attributed to weighting was predictably less in the middle of the scale where the forms were targeted. For the particular tests and number of items deleted (and weighted) in this study there was between approximately a 5% to 20% increase in standard error at this point. Error in scores containing weighted c.r items increased more substantially in the upper end of the scale where there was a 20 to 45% reduction in precision. The magnitude of increase in error and the particular portion of the scale where it occurs are determined by the locations and amount of information contributed by the deleted c.r. items relative to those that are retained.

The greater difficulty of the c.r. items meant that the location of the deleted items would tend to fall in the upper half of the scale score range, implying the total information contributed by the remaining items would be less in this part of the scale (greater error). The weighting of the retained c.r. items, though tending to be of the same difficulty as the deleted c.r. items, doesn't produce as much information as that contributed by the deleted items. Each variance of a weighted item in the denominator of equation 4 is multiplied by the square of the applied weight. The sum of the item variances subsequently increase faster than the square of the sum of derivatives $\{P'_{ik}(\theta)\}$ for the weighted (and unweighted) items in the numerator, resulting in less information and hence greater error.

Summed ratings, which increases the relative contribution of c.r. items to the total test by adding scoring levels beyond those specified by the rubrics rather than multiplying a response by a factor, results in total scores with standard errors less than that of the tcc component weighted scores throughout most of the score range. *Summed* ratings result in greater error than *Baseline* because the amount of information accrued from the additional levels is not twice the amount contributed by an e.r. item in the tests employed in the study. It is conceivable, if not likely, that there may be some c.r. items in other tests from which information gains of this magnitude could be attained.

Weighting from one through five student constructed responses by summing or multiplying by a factor of two (*Crx2* and *Erx2* analyses) resulted in differences in scale scores that most frequently (80%) differed by no more than 13 scale score points from those obtained when additional items were administered. A small difference in the difficulties of deleted and retained c.r. items contributed to slightly larger differences for the Writing Grade 8 test. Quadrupling the c.r. weighting substantially increased the mean differences and came close to doubling the 10th and 90^{th} percentile scale score differences.

The greater unreliability in the scoring of the Writing as opposed to the Math c.r. items likely contributed to the greater differences for this content area. The potential to increase score precision by improved rubrics and scoring, along with the magnitude of error at important portions of the scale, such as

cutscores, should be addressed prior to weighting c.r. items.

There are several other validity-related considerations that need to inform a decision to weight. The dimensionality assessments of the *Baseline* forms indicated one test - Math Grade 5 - was not unidimensional, having a significant second dimension defined by the c.r. items. If the multidimensionality is due to an enduring domain attribute or proficiency rather than a characteristic unique to the particular sampled c.r. items there is a potential impact on important psychometric functions such as form equating. Tcc component weighting may pose less of a problem than *Summed* Ratings under these circumstances because of its implementation "outside" of the IRT scale.

The effects of weighting on score precision and the threat that multidimensionality impairs the accuracy of the standard errors must be evaluated in light of the purpose of testing. Higher stakes testing, with the greater consequences for the student that attend score interpretation, requires at the very least a documentation of the sources and magnitude of disturbances to model-based reliability estimates as a prerequisite to a valuation. It would also seem to require a demonstration of how greater validity is obtained by increasing the representation of c.r. items through weighting rather than the number of items. Pursuant to that goal would be the presentation of evidence that the assessment of content or processes are sufficiently important to justify weighting rather than an increase in testing time.

References

Burket, G.R. (1991; 1995). *PARDUX*. Monterey, CA: CTB/McGraw-Hill.

Fitzpatrick, A.R. (1990). Status report on the results of preliminary analyses of dichotomous and multi-level items using the PARMATE (PARDUX) program. Unpublished manuscript.

Ito, K, & Sykes, R.C. (2000). *An evaluation of "intentional" Weighting of extended-response or constructed-response items in tests with mixed item types*. Paper presented at the annual National Conference on Large Scale Assessment, Snowbird, Ut.

Li, H. & Stout, W. (1995). *A version of Dimtest to assess latent trait unidimensionality for mixed polytomous and dichotomous item response data*. Paper presented at the 1995 NCME Annual Meeting, April 20, 1995.

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

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

Muraki, E. (1992). A generalized partial credit model: Applicationof an EM algorithm. *Applied Psychological Measurement, 16*, 159-176.

Thissen, D. (1986). *MULTILOG: Multiple categorical item analysis and test scoring, Version 5*. Mooresville, IN: Scientific Software.

Traub, R. E. (1993). On the equivalence of the traits assessed by multiple-choice and constructed-response tests. in R. E. Bennett & W. C. Ward (Ed.), *Construction versus choice in cognitive measurement* (pp. 29-44). Hillsdale, NJ: Lawrence Erlbaum Associates.

Wainer, H., & Thissen, D. (1993). Combining multiple-choice and constructed-response test scores: Toward a Marxist Theory of test construction. *Applied Measurement in Education, 6*, 103-118.

Table 1
Retained and Deleted C.R. Item Sets

CRx2											
Math 5				Math 8				Writing 8			
Retained		Deleted		Retained		Deleted		Retained		Deleted	
Item	P-value	Item	P-value	Item	P-value	Item	P-value	Item	P-value	Item	P-value
6	0.283	9	0.032	4	0.122	10	0.047	5	0.563	13	0.471
26	0.021	15	0.059	18	0.148	15	0.186	10	0.245	24	0.364
28	0.414	20	0.335	27	0.073	21	0.227	18	0.766	27	0.649
38	0.106	33	0.185	41	0.072	32	0.145	32	0.521	33	0.518
42	0.095	35	0.334	42	0.300	36	0.094				
Mean	0.184		0.189		0.143		0.140		0.524		0.501
SD	0.161		0.145		0.094		0.072		0.214		0.118

ERx2 and Summed							
Writing 3				Writing 8			
Retained		Deleted		Retained		Deleted	
Item	P-value	Item	P-value	Item	P-value	Item	P-value
34	0.490	33	0.471	32	0.521	33	0.518

Table 2
Raw Score Descriptive Statistics

Content	Grade	N	Points Possible	Baseline							Weighted								
				Item Test p-value		Item Test Correlation				Stratified Alpha	CRx2		Stratified Alpha	ERx2		Stratified Alpha	Summed		Stratified Alpha
				Mean	SD	Mean	SD	Mean	SD		Mean	SD		Mean	SD		Mean	SD	
Math	5	2385	55	0.455	0.222	0.380	0.090	21.91	8.81	0.871	21.86	8.76	0.831	-	-	-	-	-	-
Math	8	2748	55	0.375	0.174	0.408	0.112	18.30	9.50	0.891	18.33	9.55	0.860	-	-	-	-	-	-
Writing	3	2466	47	0.686	0.126	0.490	0.072	29.95	8.54	0.920 [1]	-	-	-	30.06	8.64	0.921	29.70	8.59	0.921
Writing	8	3288	49	0.618	0.171	0.426	0.088	28.63	7.87	0.868 [2]	28.83	8.02	0.857	28.65	8.06	0.894	28.30	8.00	0.892

1 No student obtained a perfect score of six on the first Writing prompt (Item # 33).

2 Only three students obtained a perfect score of six on the second Writing prompt (Item #33).

27

28

Table 3
Poly-Dimtest Significance Tests for the Hypothesis of Unidimensionality

		Baseline		
Content	Grade	No. Items	T	*p*-value
Math	5	45	1.779	0.038 *
	8	45	-1.070	0.858
Writing	3	34	0.625	0.266
	8	33	-0.849	0.802

* $p < .05$

Table 4

Item Total Information for the Baseline Forms

	Math				Writing			
	Grade 5		Grade 8		Grade 3		Grade 8	
Item No.	Item Location #	Total Info.*	Item Location #	Total Info.*	Item Location	Total Info.*	Item Location #	Total Info.*
1	550	0.016	533	0.016	460	0.019	490	0.014
2	563	0.012	568	0.015	461	0.015	489	0.008
3	487	0.019	548	0.020	453	0.025	438	0.014
4	388	0.006	586	0.033 [1]	522	0.021	623	0.008
5	516	0.015	567	0.024	474	0.019	476	0.026 [1]
6	573	0.037 [1]	586	0.020	494	0.027	462	0.016
7	470	0.017	508	0.016	471	0.027	454	0.015
8	458	0.024	610	0.010	466	0.022	498	0.015
9	617	0.059 [1]	554	0.027	464	0.034	445	0.017
10	558	0.016	647	0.028 [1]	547	0.013	640	0.018 [1]
11	552	0.010	607	0.018	517	0.028	454	0.024
12	576	0.027	551	0.032	506	0.033	523	0.015
13	592	0.018	497	0.016	488	0.021	507	0.028 [1]
14	492	0.026	560	0.009	506	0.024	398	0.016
15	607	0.045 [1]	560	0.045 [1]	510	0.020	576	0.010
16	438	0.014	541	0.009	498	0.027	484	0.020
17	571	0.014	575	0.007	498	0.028	524	0.021
18	584	0.014	585	0.031 [1]	486	0.020	439	0.026 [1]
19	426	0.015	593	0.003	534	0.014	514	0.016
20	529	0.033 [1]	547	0.042	574	0.021	497	0.022
21	558	0.014	557	0.034 [1]	487	0.029 [1]	529	0.027
22	575	0.014	502	0.018	470	0.017	554	0.021
23	545	0.010	531	0.019	458	0.021	471	0.019
24	602	0.020	583	0.023	453	0.023	540	0.025 [1]
25	557	0.008	614	0.017	459	0.029 [1]	476	0.014
26	634	0.044 [1]	586	0.014	559	0.026	594	0.012
27	557	0.016	585	0.084 [1]	519	0.023	455	0.023 [1]
28	514	0.034 [1]	549	0.017	442	0.031	443	0.018
29	478	0.018	547	0.027	477	0.018	479	0.022
30	561	0.017	534	0.039	499	0.024	453	0.031
31	520	0.018	547	0.052	429	0.029 [1]	508	0.017
32	523	0.028	567	0.062 [1]	417	0.021	411	0.076 [2]
33	572	0.028 [1]	489	0.013	404	0.060 [3]	400	0.061 [2]
34	505	0.017	509	0.014	405	0.063 [2]		
35	539	0.032 [1]	571	0.032				
36	584	0.037	593	0.035 [1]				
37	557	0.028	484	0.016				
38	601	0.044 [1]	566	0.033				
39	539	0.025	550	0.030				
40	537	0.012	560	0.039				
41	570	0.027	584	0.069 [1]				
42	601	0.032 [1]	546	0.029 [1]				
43	525	0.015	551	0.030				
44	590	0.014	562	0.017				
45	425	0.013	531	0.011				
Mean m.c.		0.017		0.021		0.023		0.017
SD		0.007		0.011		0.005		0.006
Mean 2-point c.r.		0.039		0.045		0.029		0.024
SD		0.009		0.020		0.000		0.004
Mean 6-point c.r.		-		-		0.062		0.068
SD		-		-		0.002		0.010

*Area under the information function

Point of maximum information

[1] Two-point CR items

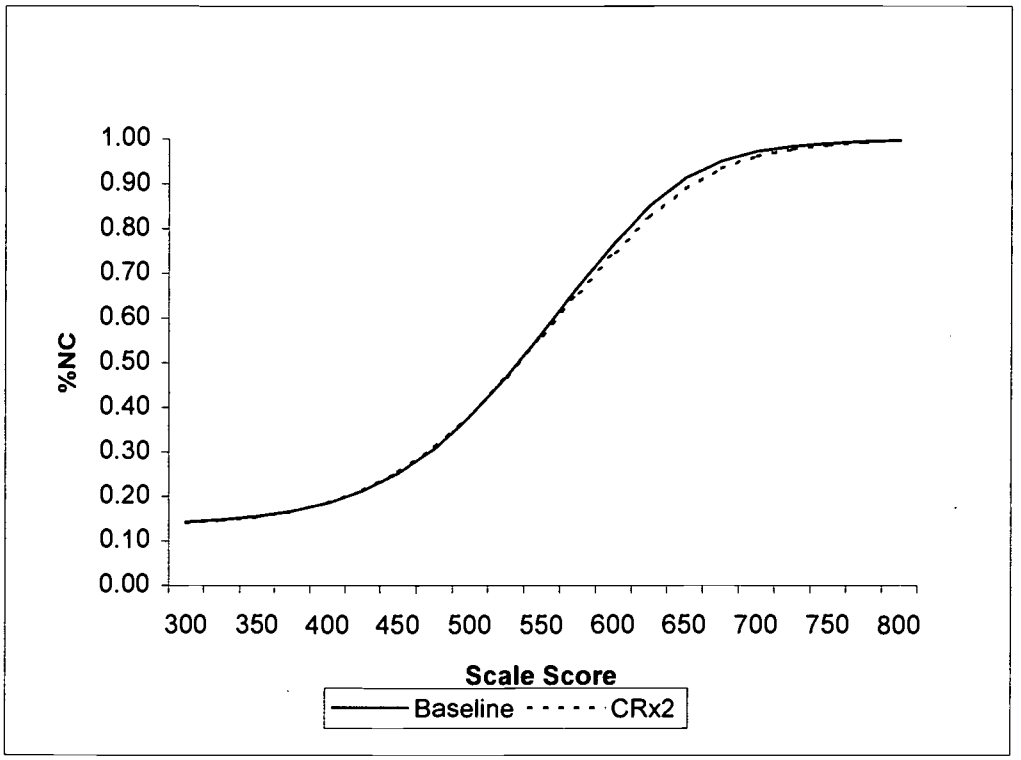
[2] Six Point Writing Prompt

[3] Writing prompt with a maximum score of 5 after collapsing one level

Table 5
Scale Score Comparisons

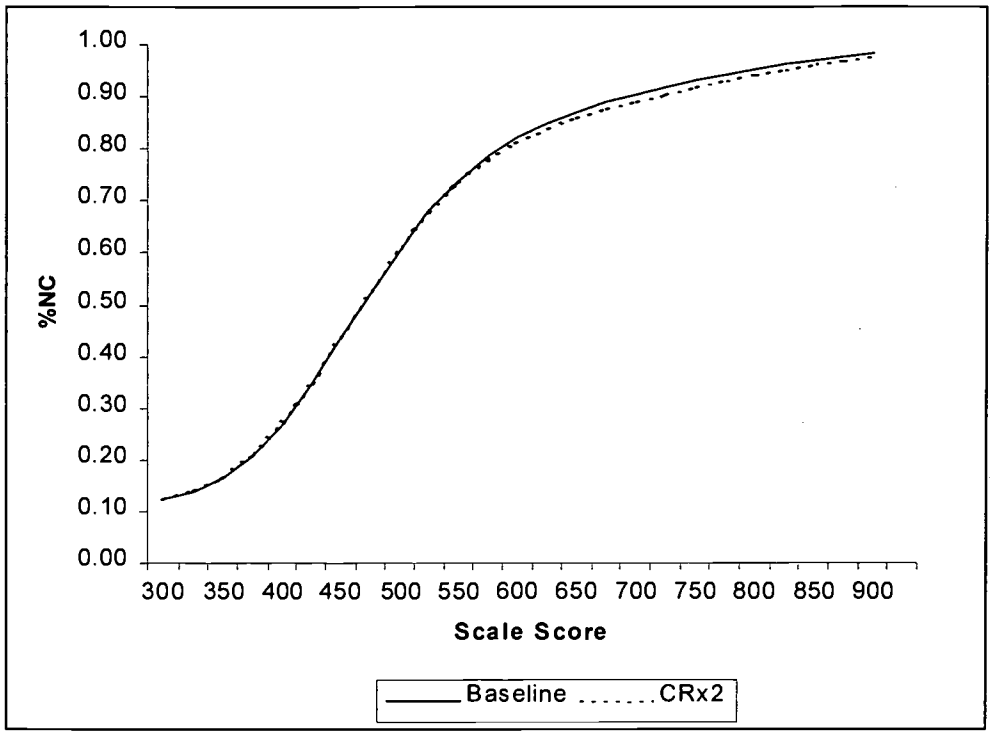
			Scale Scores												Scale Score Differences															
			Baseline		Summed		CRx2		CRx4		ERX2		ERX4		Summed - Baseline				CRx2 - Baseline				CRx4 - Baseline				ERx2 - Baseline				ERx4 - Baseline			
																	%ile				%ile				%ile				%ile			
Content	Grade	N	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	10	90	Mean	SD	10	90	Mean	SD	10	90	Mean	SD	10	90	Mean	SD	10	90
Math	5	2385	492.69	63.15	-	-	492.44	64.07	491.63	65.53	-	-	-	-	-	-	-	-	-0.25	11.20	-13	13	-1.06	18.02	-21	20	-	-	-	-	-	-	-	-
Math	8	2748	485.49	75.20	-	-	485.61	72.60	484.13	73.32	-	-	-	-	-	-	-	-	0.11	14.71	-11	12	-1.36	22.25	-24	24	-	-	-	-	-	-	-	-
Writing	3	2466	503.65	61.15	502.71	61.25	-	-	-	-	503.17	65.06	504.92	70.39	-0.94	11.93	-9	7	-	-	-	-	-	-	-	-	-0.48	14.63	-10	9	1.27	24.09	-18	19
Writing	8	3288	499.69	58.15	500.72	58.48	502.60	63.75	504.98	70.09	499.90	60.69	501.43	65.25	1.03	8.44	-8	10	2.09	17.45	-16	21	5.29	27.35	-27	37	0.21	9.24	-8	9	1.74	18.74	-17	18

Figure 1
Test Characteristic for the Math Grade 5 Forms



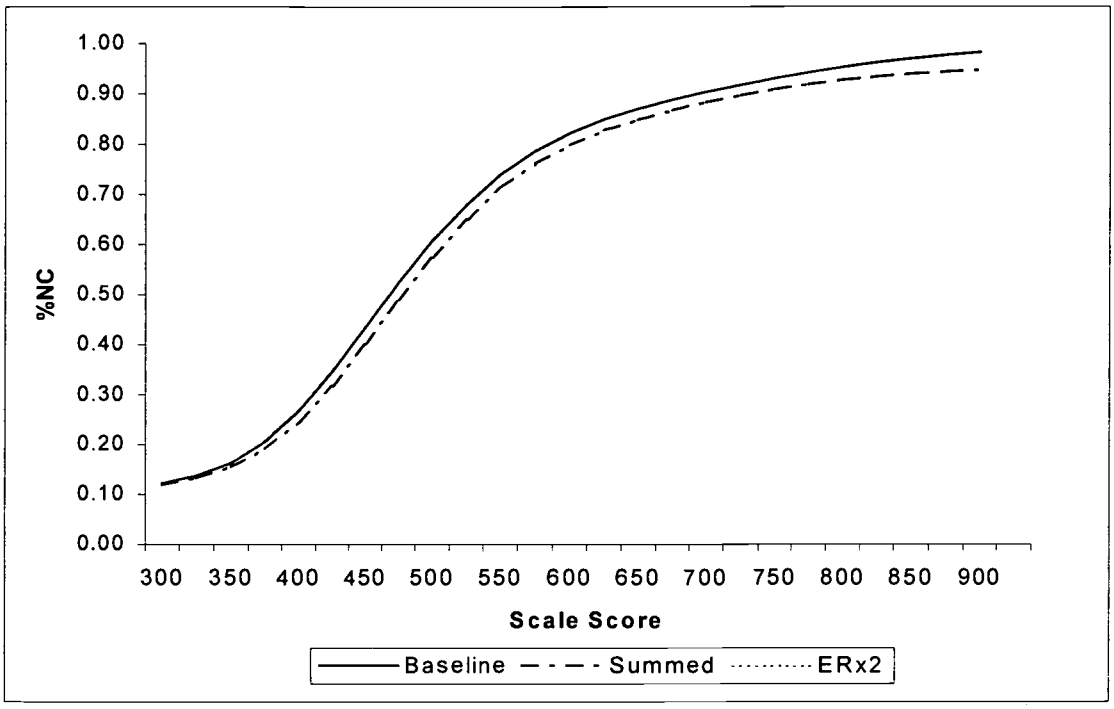
	Baseline		CRx2	
SS	RS	SE	RS	SE
300	7.88	279.79	7.77	270.02
325	8.15	184.73	8.06	178.35
350	8.57	121.67	8.49	117.88
375	9.22	81.06	9.17	79.21
400	10.20	55.21	10.21	54.84
425	11.68	38.66	11.76	39.41
450	13.86	28.13	14.02	29.74
475	16.91	21.74	17.14	24.03
500	20.92	18.05	21.11	20.84
525	25.76	15.91	25.76	18.96
550	31.18	14.64	30.87	17.87
575	36.81	14.23	36.13	18.20
600	42.17	14.90	41.07	20.16
625	46.80	16.72	45.41	22.90
650	50.21	20.78	48.94	26.95
675	52.30	27.92	51.40	34.12
700	53.46	38.38	52.89	45.09
725	54.10	52.16	53.74	59.79
750	54.46	69.59	54.24	78.37
775	54.67	91.38	54.53	101.51
800	54.80	118.62	54.71	130.35

Figure 2
Test Characteristic Curves for the Writing Grade 8 Forms:
CRx2 and Baseline



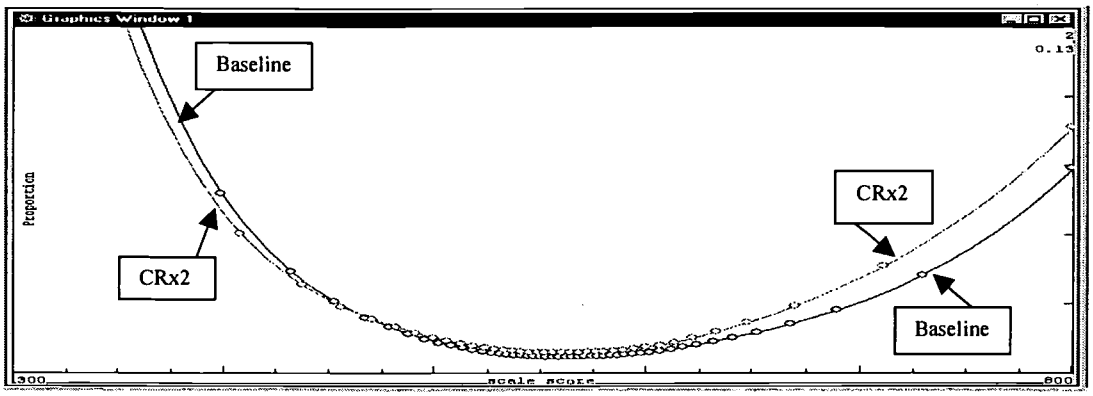
	Baseline		CRx2	
SS	RS	SE	RS	SE
300	5.98	98.80	6.05	107.17
325	6.73	63.63	6.84	73.32
350	8.00	42.77	8.15	52.15
375	10.06	30.91	10.22	39.33
400	13.08	24.43	13.23	31.67
425	16.94	20.89	17.06	26.92
450	21.28	19.08	21.40	24.16
475	25.65	18.71	25.75	23.32
500	29.71	19.28	29.76	23.81
525	33.27	20.67	33.22	25.59
550	36.23	23.29	36.04	29.49
575	38.54	27.43	38.20	36.10
600	40.29	32.65	39.81	44.57
625	41.62	38.24	41.04	53.33
650	42.67	43.65	42.04	61.40
675	43.54	48.48	42.89	68.31
700	44.29	52.44	43.65	73.89
725	44.96	55.61	44.34	78.31
750	45.58	58.38	45.00	82.13
775	46.15	61.33	45.61	86.07
800	46.67	64.99	46.17	90.79
825	47.13	69.81	46.67	96.84
850	47.52	76.18	47.11	104.61
875	47.85	84.40	47.49	114.38
900	48.12	94.74	47.81	126.38

Figure 3
Test Characteristic Curves for the Writing Grade 8 Forms:
ERx2, Summed and Baseline



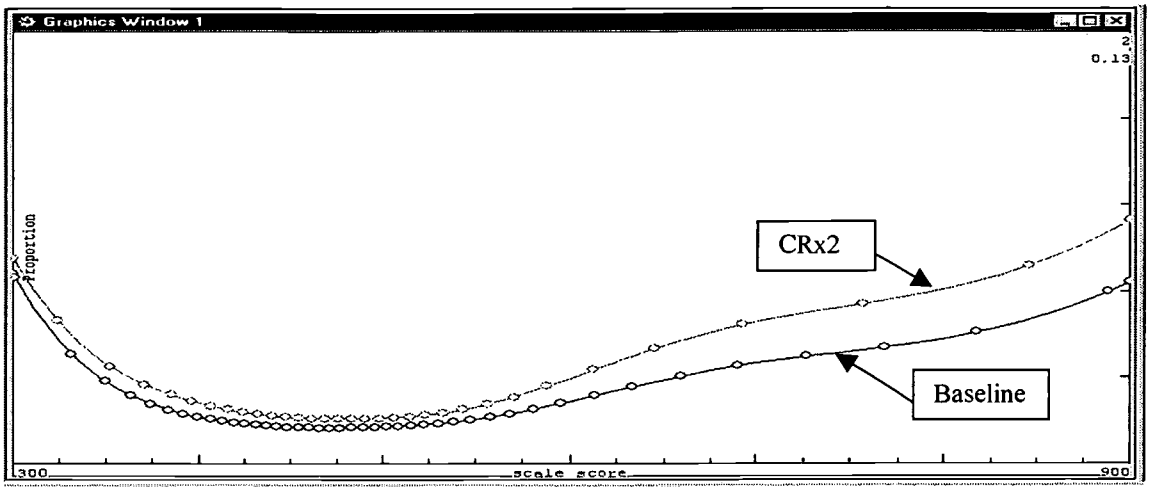
	Baseline		ERx2		Summed	
SS	RS	SE	RS	SE	RS	SE
300	5.98	98.80	5.87	93.27	5.83	109.25
325	6.73	63.63	6.72	63.71	6.50	72.31
350	8.00	42.77	8.11	45.89	7.61	49.65
375	10.06	30.91	10.28	35.28	9.38	36.13
400	13.08	24.43	13.34	28.92	12.02	28.16
425	16.94	20.89	17.15	24.77	15.53	23.33
450	21.28	19.08	21.38	22.09	19.68	20.55
475	25.65	18.71	25.65	20.95	24.01	19.56
500	29.71	19.28	29.67	21.01	28.13	19.74
525	33.27	20.67	33.25	22.22	31.82	20.96
550	36.23	23.29	36.24	25.09	34.90	23.69
575	38.54	27.43	38.57	29.95	37.31	28.23
600	40.29	32.65	40.34	36.37	39.12	34.07
625	41.62	38.24	41.68	43.65	40.51	40.46
650	42.67	43.65	42.74	51.31	41.61	46.89
675	43.54	48.48	43.60	58.88	42.51	53.06
700	44.29	52.44	44.35	65.82	43.28	58.88
725	44.96	55.61	45.01	71.88	43.95	64.61
750	45.58	58.38	45.62	77.25	44.52	70.78
775	46.15	61.33	46.17	82.52	45.01	77.96
800	46.67	64.99	46.67	88.38	45.41	86.55
825	47.13	69.81	47.11	95.43	45.74	96.79
850	47.52	76.18	47.49	104.13	46.01	108.82
875	47.85	84.40	47.81	114.84	46.22	122.70
900	48.12	94.74	48.07	127.84	46.38	138.53

Figure 4
Standard Error Curves for the CR Analyses of Math Grade 8



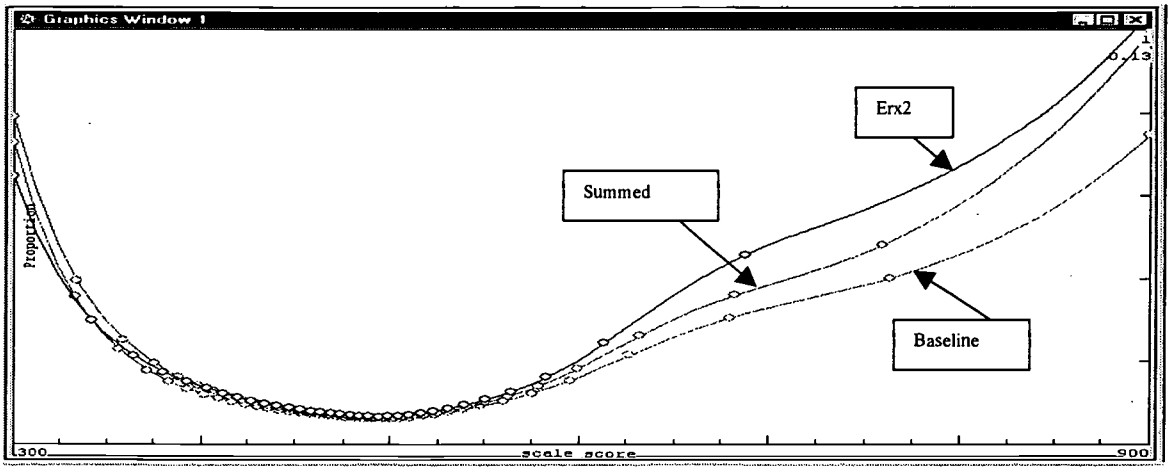
	Baseline		CRx2	
RS	SS	SE	SS	SE
0	300	198	300	194
1	300	198	300	194
2	300	198	300	194
3	300	198	300	194
4	300	198	300	194
5	300	198	300	194
6	300	198	300	194
7	300	198	300	194
8	308	190	339	155
9	399	99	402	92
10	432	66	431	63
11	453	47	450	47
12	467	36	464	39
13	479	30	475	33
14	488	25	485	29
15	496	22	493	26
16	503	20	500	23
17	508	18	506	21
18	514	16	511	19
19	519	15	517	18
20	523	14	521	17
21	527	13	525	16
22	531	13	529	15
23	534	12	533	15
24	538	12	537	14
25	541	11	540	14
26	544	11	544	14
27	547	11	547	14
28	550	11	550	13
29	553	11	553	13
30	556	11	556	13
31	559	11	559	13
32	562	11	562	13
33	565	11	566	14
34	568	11	569	14
35	571	11	572	14
36	574	11	575	14
37	577	11	578	14
38	580	11	581	14
39	583	11	584	15
40	587	12	587	15
41	590	12	591	15
42	594	13	594	16
43	598	13	598	16
44	602	14	602	17
45	607	15	606	18
46	612	16	611	19
47	617	17	616	21
48	624	18	622	23
49	632	21	629	25
50	641	23	638	29
51	652	27	649	33
52	668	32	663	40
53	689	41	685	52
54	729	64	726	81
55	800	133	800	155

Figure 5
Standard Error Curves for the CR Analyses of Writing Grade 8



	Baseline		CRx2	
RS	SS	SE	SS	SE
0	300	89	300	98
1	300	89	300	98
2	300	89	300	98
3	300	89	300	98
4	300	89	300	98
5	300	89	300	98
6	300	89	300	98
7	330	59	329	69
8	349	44	348	54
9	363	36	362	45
10	374	32	373	40
11	383	29	382	37
12	392	26	391	34
13	399	25	398	32
14	406	23	405	30
15	413	22	412	29
16	419	22	418	28
17	425	21	425	27
18	431	20	431	26
19	437	20	436	25
20	443	20	442	25
21	448	19	448	24
22	454	19	453	24
23	460	19	459	24
24	465	19	465	23
25	471	19	471	23
26	477	19	476	23
27	483	19	482	23
28	489	19	489	23
29	496	19	495	24
30	502	19	502	24
31	509	20	509	24
32	516	20	516	25
33	523	20	523	25
34	531	21	531	26
35	539	22	540	28
36	548	23	550	29
37	557	24	560	32
38	568	26	572	35
39	581	28	587	40
40	595	32	604	46
41	612	35	624	53
42	633	40	649	61
43	659	46	679	69
44	690	51	712	76
45	726	56	750	82
46	768	61	792	89
47	818	68	843	102
48	888	89	900	126
49	900	95	900	126

Figure 6
Standard Error Curves for the ER Analyses of Writing Grade 3



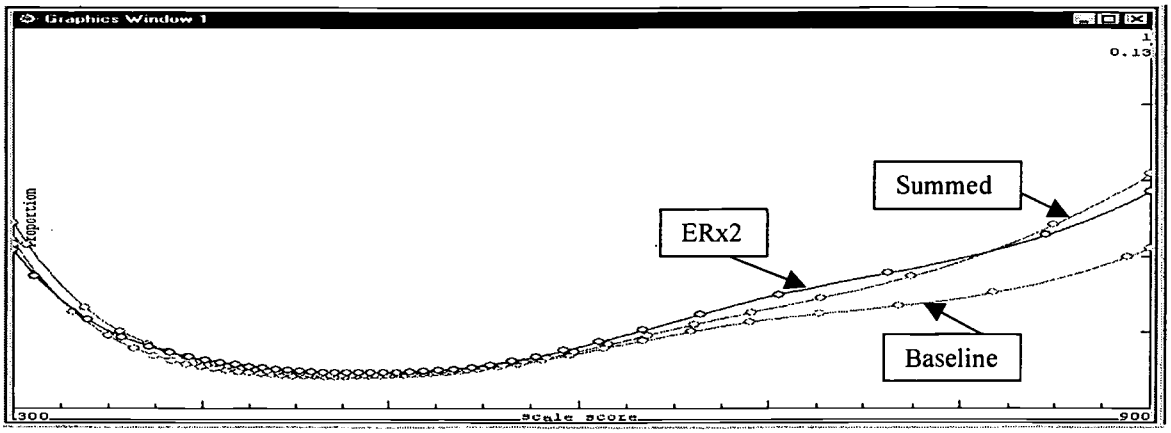
	Baseline[1]		Summed[2]		ERx2[3]	
RS	SS	SE	SS	SE	SS	SE
0	300	108	300	116	300	110
1	300	108	300	116	300	110
2	300	108	300	116	300	110
3	300	108	300	116	300	110
4	300	108	300	116	300	110
5	300	108	300	116	300	110
6	300	108	300	116	300	110
7	300	108	300	116	300	110
8	332	76	334	81	329	80
9	356	52	359	57	352	58
10	372	41	376	44	367	47
11	384	34	388	37	379	40
12	394	30	398	32	389	35
13	402	27	407	29	398	32
14	410	25	415	26	406	30
15	417	23	423	24	414	27
16	424	22	429	22	420	25
17	430	20	436	21	427	24
18	436	19	442	20	433	22
19	442	18	447	19	439	21
20	447	18	453	18	444	20
21	453	17	458	17	450	19
22	458	16	463	17	455	19
23	463	16	468	16	460	18
24	467	15	473	16	465	17
25	472	15	478	15	470	17
26	477	14	482	15	475	16
27	481	14	487	15	480	16
28	486	14	492	15	484	16
29	491	14	497	15	489	15
30	496	14	502	15	494	15
31	501	14	507	15	499	15
32	506	14	513	15	504	16
33	511	15	518	16	510	16
34	517	15	525	17	515	17
35	523	16	532	19	522	18
36	531	17	541	20	529	19
37	539	19	551	23	537	21
38	549	21	563	26	547	23
39	560	24	578	31	558	26
40	575	28	599	41	573	32
41	595	35	632	59	592	41
42	627	48	682	81	623	62
43	679	68	759	108	677	97
44	763	90	900	224	767	135
45	900	168			900	234

[1] A maximum of 45, rather than 47 points is possible because of the collapse of the uppermost category for each Writing prompt (0 and 1 student obtained a perfect score).

[2] A maximum of 44, rather than 47 points is possible because of the absence of students in the three highest categories for the Summed Writing rating prompt.

[3] A maximum of 45, rather than 47 points is possible because of the collapse of the uppermost category in the doubled Writing prompt.

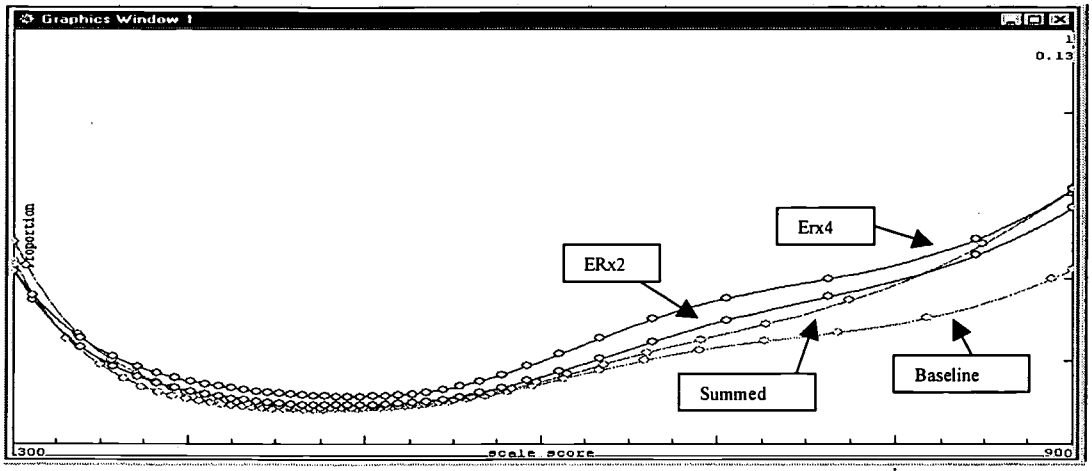
Figure 7
Standard Error Curves for the ER Analyses of Writing Grade 8



	Baseline		Summed[1]		ERx2	
RS	SS	SE	SS	SE	SS	SE
0	300	89	300	97	300	90
1	300	89	300	97	300	90
2	300	89	300	97	300	90
3	300	89	300	97	300	90
4	300	89	300	97	300	90
5	300	89	300	97	300	90
6	300	89	308	90	305	85
7	330	59	338	59	331	58
8	349	44	357	45	348	47
9	363	36	371	38	362	40
10	374	32	382	34	372	36
11	383	29	391	30	382	33
12	392	26	400	28	390	31
13	399	25	408	26	398	29
14	406	23	415	25	405	28
15	413	22	422	24	411	27
16	419	22	428	23	418	26
17	425	21	434	22	424	25
18	431	20	440	21	430	24
19	437	20	446	21	436	23
20	443	20	452	20	442	23
21	448	19	458	20	448	22
22	454	19	463	20	454	22
23	460	19	469	20	459	21
24	465	19	475	20	465	21
25	471	19	481	20	471	21
26	477	19	487	20	477	21
27	483	19	493	20	483	21
28	489	19	499	20	489	21
29	496	19	506	20	496	21
30	502	19	512	20	502	21
31	509	20	519	21	509	21
32	516	20	526	21	516	22
33	523	20	534	22	523	22
34	531	21	542	23	531	23
35	539	22	551	24	539	24
36	548	23	561	25	548	25
37	557	24	571	27	557	26
38	568	26	584	30	568	28
39	581	28	598	34	580	31
40	595	32	615	38	595	35
41	612	35	636	43	612	40
42	633	40	660	49	632	46
43	659	46	690	57	657	54
44	690	51	727	65	688	63
45	726	56	774	78	724	72
46	768	61	849	108	767	81
47	818	68	900	139	818	93
48	888	89			893	124
49	900	95			900	128

[1]A maximum of 47, rather than 49 points is possible because of the absence of students in the two highest categories of the Summed Rating Writing prompt.

Figure 8
Writing Grade 8: Multiple Weighting Types

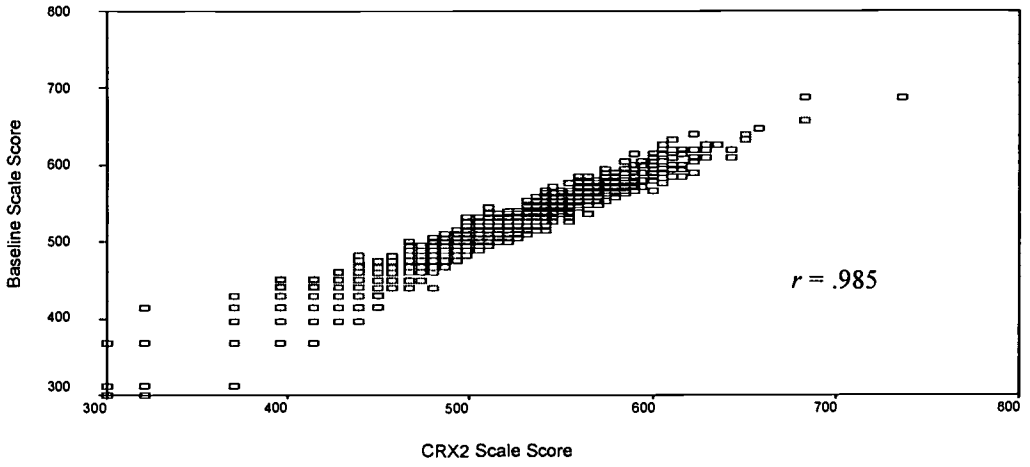


	Baseline		Summed[1]		ERx2		ERx4	
RS	SS	SE	SS	SE	SS	SE	SS	SE
0	300	89	300	97	300	90	300	93
1	300	89	300	97	300	90	300	93
2	300	89	300	97	300	90	300	93
3	300	89	300	97	300	90	300	93
4	300	89	300	97	300	90	300	93
5	300	89	300	97	300	90	301	92
6	300	89	308	90	305	85	328	65
7	330	59	338	59	331	58	345	54
8	349	44	357	45	348	47	357	47
9	363	36	371	38	362	40	368	43
10	374	32	382	34	372	36	377	40
11	383	29	391	30	382	33	384	38
12	392	26	400	28	390	31	392	36
13	399	25	408	26	398	29	398	35
14	406	23	415	25	405	28	405	34
15	413	22	422	24	411	27	411	32
16	419	22	428	23	418	26	417	32
17	425	21	434	22	424	25	423	31
18	431	20	440	21	430	24	429	30
19	437	20	446	21	436	23	435	29
20	443	20	452	20	442	23	441	28
21	448	19	458	20	448	22	447	28
22	454	19	463	20	454	22	453	27
23	460	19	469	20	459	21	459	27
24	465	19	475	20	465	21	465	26
25	471	19	481	20	471	21	472	26
26	477	19	487	20	477	21	478	26
27	483	19	493	20	483	21	485	26
28	489	19	499	20	489	21	492	26
29	496	19	506	20	496	21	499	26
30	502	19	512	20	502	21	507	26
31	509	20	519	21	509	21	514	26
32	516	20	526	21	516	22	523	27
33	523	20	534	22	523	22	531	28
34	531	21	542	23	531	23	541	29
35	539	22	551	24	539	24	551	31
36	548	23	561	25	548	25	562	33
37	557	24	571	27	557	26	575	37
38	568	26	584	30	568	28	589	41
39	581	28	598	34	580	31	606	47
40	595	32	615	38	595	35	626	55
41	612	35	636	43	612	40	650	63
42	633	40	660	49	632	46	676	72
43	659	46	690	57	657	54	706	79
44	690	51	727	65	688	63	738	85
45	726	56	774	78	724	72	772	91
46	768	61	849	108	767	81	809	99
47	818	68	900	139	818	93	855	115
48	888	89			893	124	900	138
49	900	95			900	128	900	138

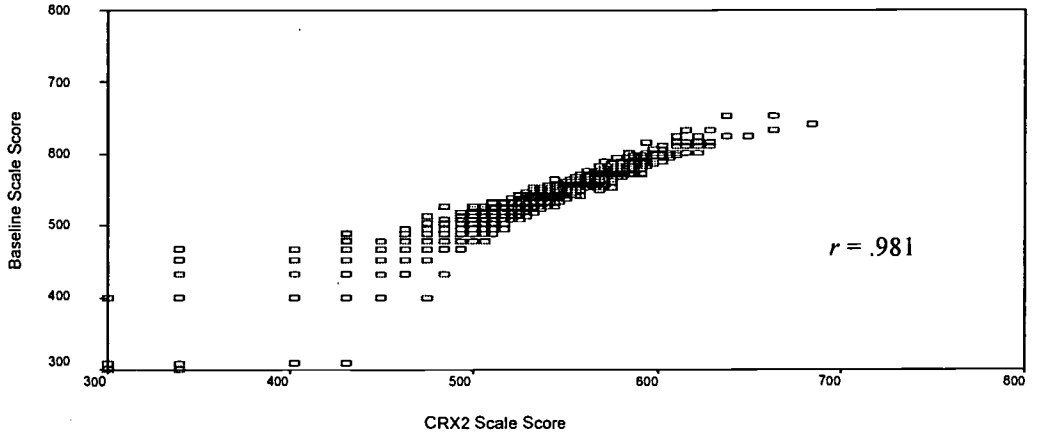
[1]A maximum of 47, rather than 49 points is possible because of the absence of students in the two highest categories of the Summed Rating Writing prompt.

Figure 9
CRx2 Weighted Scale Scores versus Baseline

Math Grade 5



Math Grade 8



Writing Grade 8

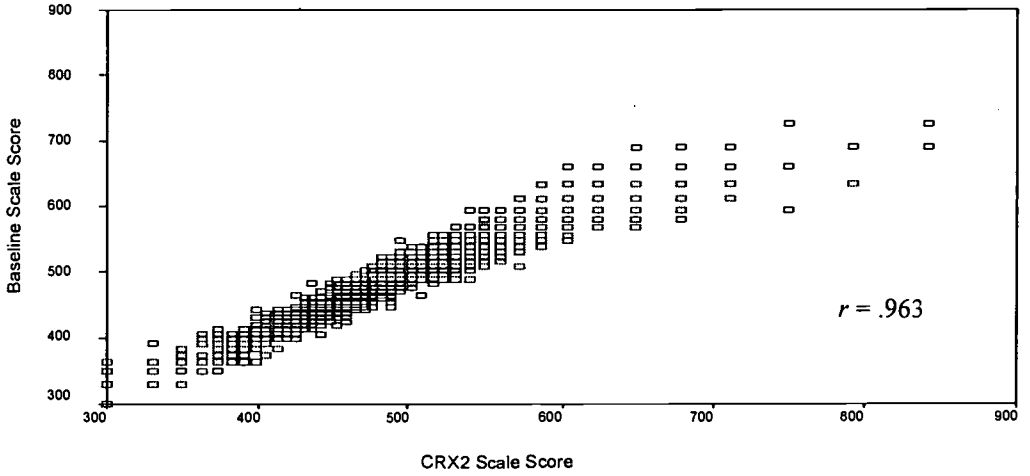
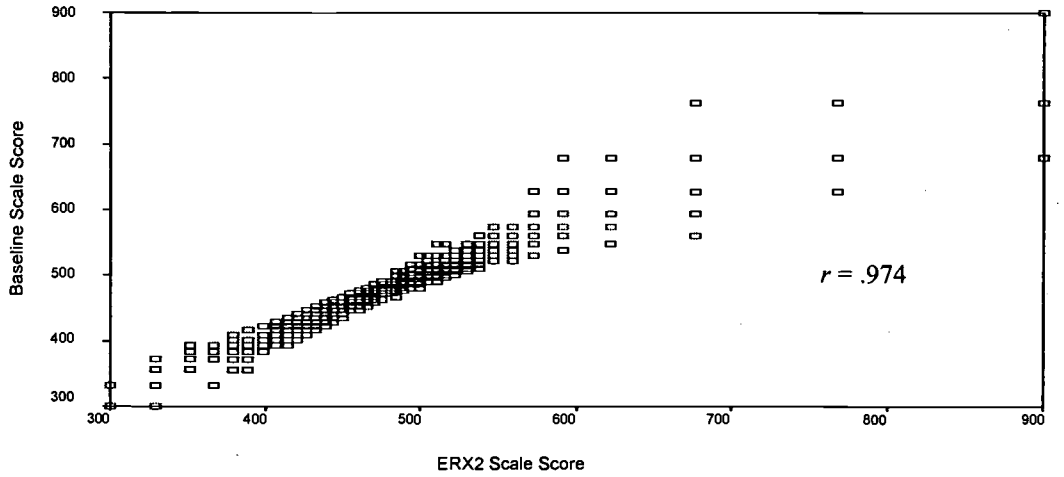
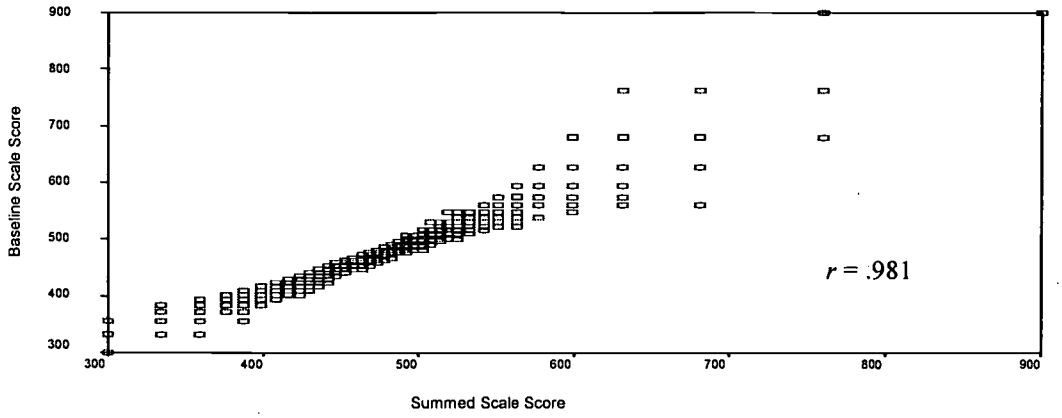
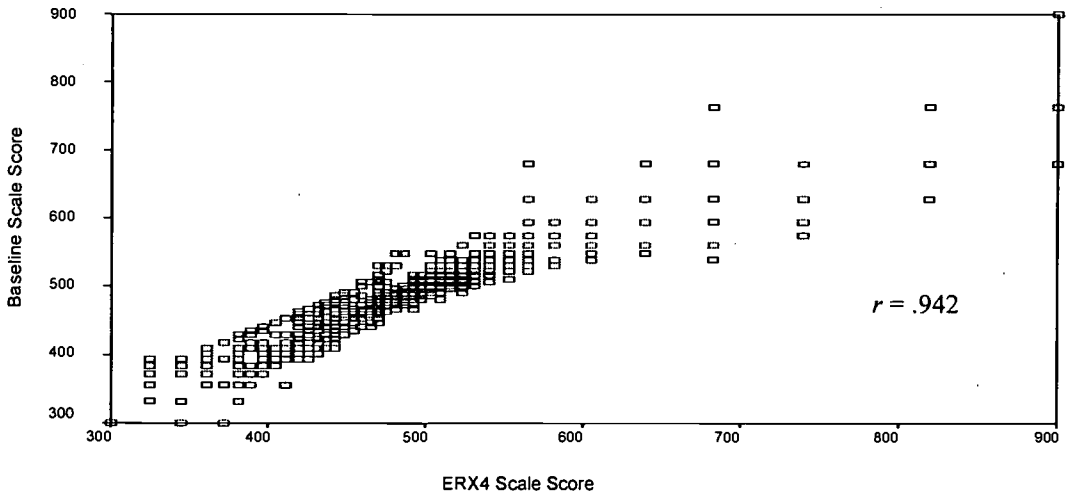


Figure 10
Writing Grade 3 Weighted versus Baseline Scale Scores







TM032803

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Determining the Representation of Constructed Response Items in Mixed-Item Format Exams

Author(s): Robert C. Sykes, Denise Truskosky, and Hillary White

Corporate Source: CTB/McGraw-Hill

Publication Date: April 2001

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

☑

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A

☐

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

☐

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, → please

Signature: Robert C. Sykes	Printed Name/Position/Title: Research Scientist III	
Organization/Address: CTB/McGraw-Hill 20 Ryan Ranch Road Monterey, CA. 93940	Telephone: (831) 393-7774	FAX:
	E-Mail Address: rsykes@ctb.com	Date: 4/18/01

(over)