Practical Assessment, Research, and Evaluation

Volume 9 *Volume 9, 2004* Article 13

2004

IRT-Linked Standard Errors of Weighted Composites

Ruth A. Childs

Susan Elgie

Tahany Gadalla

Ross Traub

Andrew P. Jaciw

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 9, Number 13, June, 2004 ISSN=1531-7714

IRT-Linked Standard Errors of Weighted Composites

Ruth A. Childs, Susan Elgie, Tahany Gadalla, & Ross Traub
Ontario Institute for Studies in Education of the University of Toronto

Andrew P. Jaciw
Stanford University

In this study, we describe an approach to calculating the standard errors of weighted scores while maintaining a link to the IRT score metric, then use the approach to compare three sets of weights. Weighting a mathematics test's multiple-choice items, short-answer items, and extended constructed-response items to achieve a ratio of 2:2:6 on the raw score metric had little effect on examinee scores or standard errors. Ratios of 3:3:4 and of 1:1:8 required more extreme weights and had a slightly larger, but still small, effect on results, increasing the standard errors. Overall, as the difference between the intended emphasis and the test's design increased, the effect of the weighting also increased.

Weighting items when forming composite scores is unavoidable. Gulliksen (1950) both begins and ends his description of weighting approaches with that observation. Whenever a score is created from more than one item or set of items, the way in which the results are combined and the psychometric properties of the components determines their contributions to the composite score. Weights can be imposed explicitly to try to control the effect of each component, but even if weights are not defined explicitly, the features of the original results, such as their ranges of values, and decisions about how to combine the results, weights their contributions implicitly.

In the simplest scenario, when the scores on individual test items are added together to create a test score, the weighting of the items is nevertheless determined by the test design. The number of possible score points for each item, the number of items of each type and measuring each part of the test's content, and the variance of the item scores determine relative contributions to the overall score.

When test items are calibrated using item response theory (IRT) models, weighting occurs as a result of the model. An advantage of these models – specifically, models in which the slope is allowed to vary across items, such as the two- and three-parameter logistic (2PL and 3PL) models and Samejima's graded response (GR) model (Samejima, 1969) – is that the items better able to discriminate among examinees will contribute more information to the final examinee score estimates and have higher weights. Lord (1980), provided an argument for using IRT to impose weights, describing how to compute the information function for a weighted composite of item scores and demonstrating that the optimal weights for dichotomously-scored items are equal to the slope parameters in the IRT model. As Thissen, Nelson, Rosa, and McLeod (2001) observe, "*Each [examinee's] response pattern is scored in a way that best uses the information about proficiency that the entire response pattern provides, assuming that the model summarizes the data accurately*" (p. 169).

Sometimes tighter controls of the contribution of items to examinee scores may be required. For example, a testing program that has set a strict policy of weighting item types (e.g., multiple-choice, short-answer, and extended constructed-response) in proportion to the amount of testing time may insist on explicit weighting of the items. For such a program, being able to state that item types are weighted in proportion to testing time (and, especially, that multiple-choice items are not making a disproportionate contribution) may override other considerations. Cohen & Jiang (2001) describe a similar situation in which a state testing program administers many multiple-choice items and only a few more costly constructed-response items. In their example, however, the test results must reflect an emphasis on the skills measured by the constructed-response items, requiring that item types be weighted to represent the desired test design, not the design dictated by cost constraints.

This study describes how to calculate the standard errors of weighted scores while maintaining a link to the IRT score metric and uses the approach to compare three sets of weights.

Approaches to Weighting

A wide variety of approaches to weighting items and sets of items have been described. For example, Gulliksen (1950), summarizing the state of the art at the time, provided formulae and rationales for basing weights on the reliabilities, error variances, or standard deviations of the test or subtest scores, on factor analysis results, or on maximizing the

reliability of the composite score. McDonald (1968) proposed a "*unified treatment of the weighting problem,*" classifying the approaches described by Gulliksen and others as special cases of a general approach. Wang & Stanley (1970) reviewed approaches to determining weights for tests, subtests, and items, and added a consideration of differential weighting of response options within items. They also reviewed the available empirical research, concluding:

> *The effectiveness of weighting depends on the number of measures to be combined, their intercorrelations, and certain characteristics of the weights. Weighting is most effective when there are but a few relatively independent variables in the composite. With a large number of positively correlated variables (such as test items), the correlation between two randomly weighted composites rapidly approaches unity.* (p. 699)

As increasing numbers of constructed-response items have joined multiple-choice items in large-scale assessments, recent studies have focused on issues related to creating weighted composites from tests or subtests consisting of different types of items. For example, Wainer & Thissen (1993) compared approaches to combining a test consisting of multiple-choice items with one consisting of constructed-response items. They summarize their results as follows:

> *[W]e ought to weight each section of a test by its value to the measurement goals of the test. This is done in the small, by item, when a test is scored with an IRT model that allows differential slopes of each item's trace line…. We added our support to those who have suggested this in the large, by test section. Is this what takes place in practice? No. Test sections are typically weighted by the amount of testing time. We have shown that such a policy will typically yield a score of lower reliability than might have been obtained with an optimal-weighting system.* (p. 113)

Rudner (2001) expresses concern that the focus on the reliability of the composite score may cause researchers to lose sight of the effect of weighting on validity. He describes a case in which a multiple-choice subtest has higher reliability, but a lower correlation with a criterion measure, than a constructed-response subtest. In such a case, determining weights may involve a trade-off between the reliability of the composite and its validity. Here the correlation between subtests is not high, so that the subtests fit Stanley & Wang's description of "*but a few relatively independent variables"* when "*weighting is most effective*" (p. 699).

Ercikan et al. (1998), similarly, describe the value of constructed-response items as increasing test validity. In their study, they compare scores and information from multiple-choice and constructed-response items calibrated together and calibrated separately using IRT models, but do not create a composite score from the results of the separate calibrations.

Related to the issue of score validity, Wilson & Wang (1995) compared the contributions of multiple-choice and performance-based items to defining the latent variable in IRT analyses and to test information, concluding that performance-based items contributed more information than multiple-choice items, but that multiple-choice items did affect the definition of the latent variable.

Ito & Sykes (2000) and Sykes, Truskosky, & White (2001) examined the effects on composite test scores of increasing the weights of extended response, constructed-response, or multiple-choice items, as compared to the effects of adding an equivalent number of items of the relevant type. The standard errors of the scores that included weighted components were higher across the latent trait scale, though the difference was greater on the lower and upper parts of the scale than in the middle.

A number of recent research studies (e.g., Ban & Lee, 2002) have investigated the psychometric properties of weighted scale scores and approaches to computing standard errors for such scores. Fewer (see, however, Cohen & Jiang, 2001; Kolen & Wang, 1998) have addressed computing the standard errors of weighted scores, while maintaining a link to the IRT score metric. This may be particularly useful if the equating is performed using the IRT metric and the cutscores are determined in that metric (see, for example, Childs, Jaciw, & Saunders, 2002).

This Study

The purpose of this study is to describe and determine the effect of an alternative scoring approach implemented in order to permit explicit weighting of items while maintaining a link to the IRT metric. In addition to describing and illustrating the approach, this study addresses the following questions for three sets of weights:

1. What are the correlations between examinees' weighted summed scores and IRT scores?
2. If these scores were reported in five categories, what percentage of examinees would be classified differently?
3. Does one approach favor high or low achievers?
4. Which approach yields a lower standard error of measurement?

METHOD

Data

The data for this study are taken from the Spring 2000 field test of the Grade 9 Assessment of Mathematics, a

standardized assessment developed by the Education Quality and Accountability Office (EQAO), Ontario's provincial testing agency. The assessment included multiple-choice items, short-answer items, and extended constructed-response items. Only examinees responding to all three item types were included in these analyses. Although different versions of the test were created to match different Grade 9 mathematics courses and languages of instruction, only one version of the test is analyzed in this study. Three parallel forms of that test version were field-tested, with approximately 1,500 observations for each form. Two of the three forms provided data for this study. These forms will be referred to as Form X and Form Y.

After a few items that were to be rewritten were dropped, Form X consisted of 23 multiple-choice items, 11 short-answer items (with a maximum total score of 22), and 23 extended constructed-response items (with a maximum total score of 96). The short-answer and extended constructed-response items were scored by trained markers using item-specific scoring guides. The number of codes in each scoring guide was determined by the complexity of the item and ranged from four to seven. Across all item types, the maximum total score was 141. The number of examinees was 1,256.

Form Y consisted of 22 multiple-choice items, 8 short-answer items (with a maximum total score of 16), and 14 extended constructed-response items (with a maximum total score of 75). Across all item types, the maximum total score was 113. The number of examinees was 1,188. The score distributions and reliabilities of the raw scores by item type and overall for both forms are provided in Table 1.

Table 1: Descriptive Statistics by Item Type

Form	Item Type	Number of Items	Score			Alpha
			Maximum	*M*	*SD*	
X (N = 1,256)	Multiple-Choice	23	23	16.12	4.04	.76
	Short-Answer	11	22	12.53	4.37	.72
	Extended Constructed-Response	23	96	88.00	15.02	.88
	All Items	57	141	120.56	21.32	.91
Y (N = 1,188)	Multiple-Choice	22	22	15.52	4.19	.80
	Short-Answer	8	16	10.07	2.37	.51
	Extended Constructed-Response	14	75	60.65	10.79	.83
	All Items	44	113	91.05	14.57	.87

Note: Because of missing data, the statistics for the item types and overall scores are based on different numbers of observations.

Previous analyses (Childs et al., 2001) investigated the dimensionality of the assessment to determine whether the assumption that the set of items being calibrated is unidimensional is violated. Confirmatory factor analyses indicated that factors defined by content strands and skill categories were very highly correlated and the models including these factors were not significant improvements over the unidimensional model. Factors defined by item types were less highly correlated; however, the correlations (between .8 and .9) were sufficiently high that the decision was made to calibrate the item types together, allowing a single latent trait to be defined by the three item types.

Computing Scores and Standard Errors

IRT scoring. Weighting in IRT is related to item information. Item information is defined as:

$$I_j(\theta) = \frac{[P_j'(\theta)]^2}{P_j(\theta)Q_j(\theta)}$$

where

$I_j(\theta)$ is the information provided by item j at ability level θ,
$P_j(\theta)$ is the probability of answering item j correctly at θ ,
$Q_j(\theta)$ is 1-P_j or the probability of answering the item incorrectly at q , and
$P_j'(\theta)$ is the first derivative of $P_j(\theta)$ and is equal to the slope of the item characteristic curve (ICC) at q .

Test information is computed by summing item information at each level of q. The standard error of a test score is related to test information as follows,

$$SE(\theta) = \frac{1}{\sqrt{I(\theta)}}$$

where $SE(\theta)$ is the standard error of estimation at that ability level. (For detailed treatments of IRT information, see Baker, 1992; Hambleton, Swaminathan, & Rogers, 1991.)

The PARSCALE program (Muraki & Bock, 1996) was used to compute Marginal Maximum Likelihood (MML) item parameter estimates for the items, applying the 3PL model to the multiple-choice items and the logistic form of the GR model to the short-answer (marked on a three-point scale) and extended constructed-response (marked on a four- to seven-point scale). PARSCALE was also used to compute examinee score estimates, based on these item parameter estimates, with omitted items ignored. Expected a Posteriori (EAP) scoring was used, in which a *N(0,1)* population prior is incorporated into the estimates. PARSCALE also yielded standard errors for the score estimates.

Alternative scoring. The testing program's policy specified that items in the test should be weighted so that the ratios of the contributions of multiple-choice, short-answer, and extended constructed-response items are exactly 2:2:6. For purposes of comparison, ratios of 3:3:4 and 1:1:8 are also used. To accommodate the weighting, item-level variances are computed in the raw score metric. Estimates were made at 61 *levels of q* equally spaced between -3 and +3. Because the item-level variances are computed at specific levels of q, they are standard errors of measurement at those levels. Weighted composites of the item scores and of the item variances at each chosen level of q are then created. Because item responses at a given level of θ are assumed to be independent, the off-diagonal elements in the correlation matrix need not be considered when combining the item variances (Lord, 1980). Finally, examinees' weighted summed scores are computed.

The specific steps are as follows:

1. Compute the conditional variance for each item at each level of q:

- Multiple-choice items:

$$s^2_{\mathrm{MC},j}(\theta) = P_j(\theta)(1 - P_j(\theta))$$

- Short-answer items:

$$s^2_{\mathrm{SA},j}(\theta) = \sum_0^m P(\mathrm{Score} \mid \theta)(\mathrm{Score}^2) - \left(\sum_0^m P(\mathrm{Score} \mid \theta)(\mathrm{Score})\right)^2$$

where the possible scores on the item range from 0 to m.

- The equation for extended constructed-response items is analogous to that for short-answer items.

2. Compute the expected item-level raw score for all items of a type at each level of q:

- Multiple-choice items (sum the item probabilities at q):

$$E(X_{\mathrm{MC}}(\theta)) = \sum_{j=1}^{n_{\mathrm{MC}}} P_{\mathrm{MC},j}(\theta)$$

where

X_{MC} is the total raw score on the multiple-choice items, and
n_{MC} is the number of multiple-choice items.

- Short-answer items (sum the probabilities of the scores within items, then sum across items):

$$E(X_{\mathrm{SA}}(\theta)) = \sum_{j=1}^{n_{\mathrm{SA}}} \sum_{k=0}^{m} kP_{\mathrm{SA},jk}(\theta)$$

where

X_{SA} is the total raw score on the short-answer items, and
n_{SA} is the number of short-answer items.

- Again, the equation for the extended constructed-response items is analogous to that for short-answer items.

3. Compute weights to be applied to the expected raw scores such that the contributions of the item types to the maximum total raw score are in the desired proportions.

- Multiple-choice items:

$$W_{\mathrm{MC}} = A_{\mathrm{MC}} \times \frac{Max_{\mathrm{Total}}}{Max_{\mathrm{MC}}}$$

where

W_{MC} is the weight for the multiple-choice items,
A_{MC} is the intended proportion of weight for the multiple-choice items (for example, for a ratio of 1:1:8, the proportion of weight for multiple-choice items would be 0.1),
Max_{Total} is the maximum score across all items of all types, and
Max_{MC} is the maximum score for the multiple-choice items.

- The equations for short-answer items and for extended constructed-response items are analogous to that for multiple-choice items.

4. At each θ level, apply these weights to the expected totals for each type of item to create a weighted linear composite of the expected raw scores:

$$X_{\mathrm{Total}}(\theta) = W_{\mathrm{MC}}E(X_{\mathrm{MC}}(\theta)) + W_{\mathrm{SA}}E(X_{\mathrm{SA}}(\theta)) + W_{\mathrm{ECR}}E(X_{\mathrm{ECR}}(\theta))$$

This weighted linear composite is analogous to a weighted test characteristic curve and so provides a means of translating between the weighted summed score metric and the IRT metric.

5. Compute a composite variance for these totals:

$$s^2_{\mathrm{Total}}(\theta) = W^2_{\mathrm{MC}} \sum_{1}^{n_{\mathrm{MC}}} s^2_{\mathrm{MC},j}(\theta) + W^2_{\mathrm{SA}} \sum_{1}^{n_{\mathrm{SA}}} s^2_{\mathrm{SA},j}(\theta) + W^2_{\mathrm{ECR}} \sum_{1}^{n_{\mathrm{ECR}}} s^2_{\mathrm{ECR},j}(\theta)$$

6. Compute the standard error as the square root of the variance at each θ level.

7. Compute examinee weighted summed score estimates:

$$Score = W_{MC}\sum_{j=1}^{n_{MC}} C_{MC,j} + W_{SA}\sum_{j=1}^{n_{SA}} C_{SA,j} + W_{ECR}\sum_{j=1}^{n_{ECR}} C_{ECR,j}$$

where

$C_{MC,\,j}$ is the raw score on multiple-choice item j, and
variables referring to the short-answer and extended constructed-response portions of the assessment are similarly defined.

Classifying Examinees

Had this been an operational administration of the test, the scores would have been divided into performance levels for reporting. For the subsequent operational administration of this test, the percentages of examinees performing at each of five performance levels were approximately, from lowest to highest level: 7%, 14%, 27%, 45%, and 6%. For the purposes of this analysis, cutpoints were set so that the percentages of examinees' scores in the five performance levels matched these percentages as closely as possible.

RESULTS

The weights computed, using Equations 9, 10, and 11, for the ratios of 1:1:8, 2:2:6, and 3:3:4 are presented in Table 2 for Forms X and Y.

Table 2: Weights, Correlations, Classification Agreement, and Standard Errors by Form and Ratio

		Weight			Correlation			
Form	Ratio	Multiple-Choice (W_{MC})	Short-Answer (W_{SA})	Extended Constructed-Response (W_{ECR})	IRT Scores	Unweighted Summed Scores	Agreement with Classifications Based on IRT Scores	Average Standard Error
X	1:1:8	0.62	0.65	1.17	.885	.974	67.3%	6.40
	2:2:6	1.23	1.29	0.88	.920	.998	73.0%	6.14
	3:3:4	1.85	1.94	0.59	.933	.996	75.6%	6.90
Y	1:1:8	0.45	0.62	1.30	.917	.987	73.1%	5.61
	2:2:6	0.90	1.24	0.97	.939	1.000	76.1%	5.12
	3:3:4	1.35	1.86	0.65	.942	.992	75.7%	5.44

Correlations. For Form X and the ratio of 2:2:6, the correlation between the examinees' IRT scores and weighted summed scores was .920. For comparison, the correlation between the unweighted summed scores and the weighted summed scores was .998, but the correlation between the unweighted summed scores and the IRT scores was .908. For the ratio 1:1:8, the correlation of the IRT scores and weighted summed scores was .885; for the ratio 3:3:4, it was .933. The correlations with the unweighted summed scores were .996 and .974, respectively. In summary, the ratio 2:2:6 yielded scores closest to the unweighted summed scores (this is not surprising, as the weights for that ratio are closest to 1; however, the 1:1:8 ratio produced results almost as highly correlated). The ratio 3:3:4 produced scores closest to the IRT scores. These correlations are presented in Table 2 to enable comparison.

Similarly, for Form Y, the correlation between the examinees' IRT scores and weighted summed scores for the 2:2:6 ratio was .938. The correlation between the unweighted summed scores and the weighted summed scores was 1.000, but the correlation between the unweighted summed scores and the IRT scores was .940. The weighted summed scores produced

by the ratio 1:1:8 were correlated .917 with the IRT scores; for the ratio 3:3:4, the correlation was .942. The correlations with the unweighted summed scores were .992 and .987, respectively.

Classification consistency For Form X and the ratio of 2:2:6, classifications were the same as those produced using the IRT scores for 73.1% of the examinees and differed by one level for 25.7%, by two levels for 1.1%, and by three levels for 0.1%. For the 1:1:8 ratio, classifications were the same for 67.3%; for 3:3:4, the classifications agreed exactly for 75.6%.

For Form Y, classifications were the same for 76.1% of the examinees in 2:2:6. Classifications differed by one level for 22.6%, and by two levels for 1.3%. For 1:1:8, classification was unchanged for 73.1%. The ratio of 3:3:4 produced exact agreement for 75.7%.

Differential accuracy. For Form X, a very small number of examinees (15; 1.2%) who were in a very low category based on their weighted summed scores moved up two or three levels from their classification based on the 2:2:6 ratio, when categorized on their IRT scores. For the 1:1:8 ratio, a few more (30; 2.4%) moved up two or more levels; for the 3:3:4 ratio, a few less changed by this amount (7; 0.6%). These changes are likely because the IRT scoring ignored omitted items, while summed scores necessarily penalize omits. However, apart from this small difference, neither scoring approach clearly favors high or low achievers. Results were very similar for Form Y.

Standard errors. The standard errors were computed as described previously. Figure 1 shows the IRT information by item type for Form X. The extended constructed-response items contribute the most information for all levels of q ; compared to the short-answer items, the multiple-choice items contribute slightly more information at the upper end of the scale. Figure 2 compares the standard errors for both approaches, with the IRT standard errors projected onto the raw score metric used by the weighted summed scores (the shape of the curve for the non-projected IRT standard errors was, naturally, the same). The weighted summed score standard errors for all three ratios are larger than those for the IRT scores for the upper and lower ends of the q scale. For a narrow range in the middle of the scale, the standard errors for the 1:1:8 and the 2:2:6 ratios are smaller than the IRT standard errors. Overall, the ratio of 1:1:8 resulted in standard errors about 5% larger than the standard errors for the 2:2:6 ratio; a ratio of 3:3:4 resulted in standard errors 10% larger. The different shapes of the curves can be seen to be related to the differences in information by item type across the range of achievement. The average standard errors by ratio and form are provided in Table 2.

Figure 1. IRT information by item type (Form X).

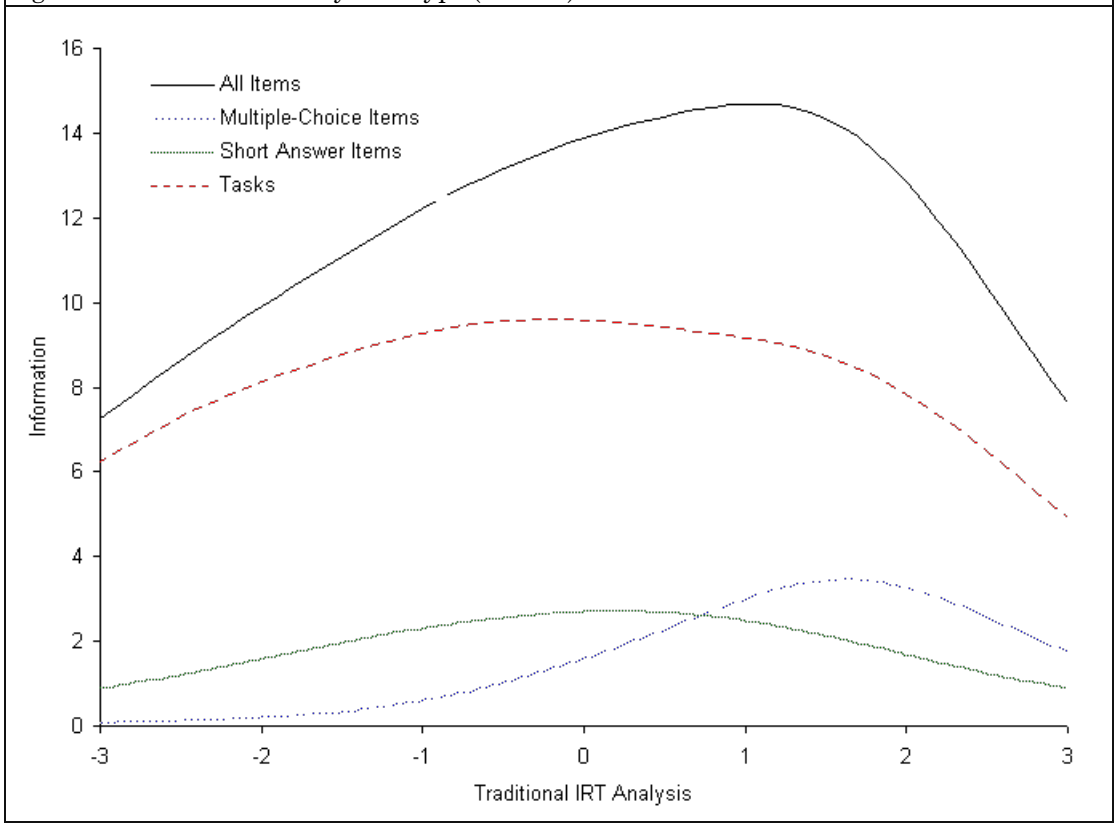
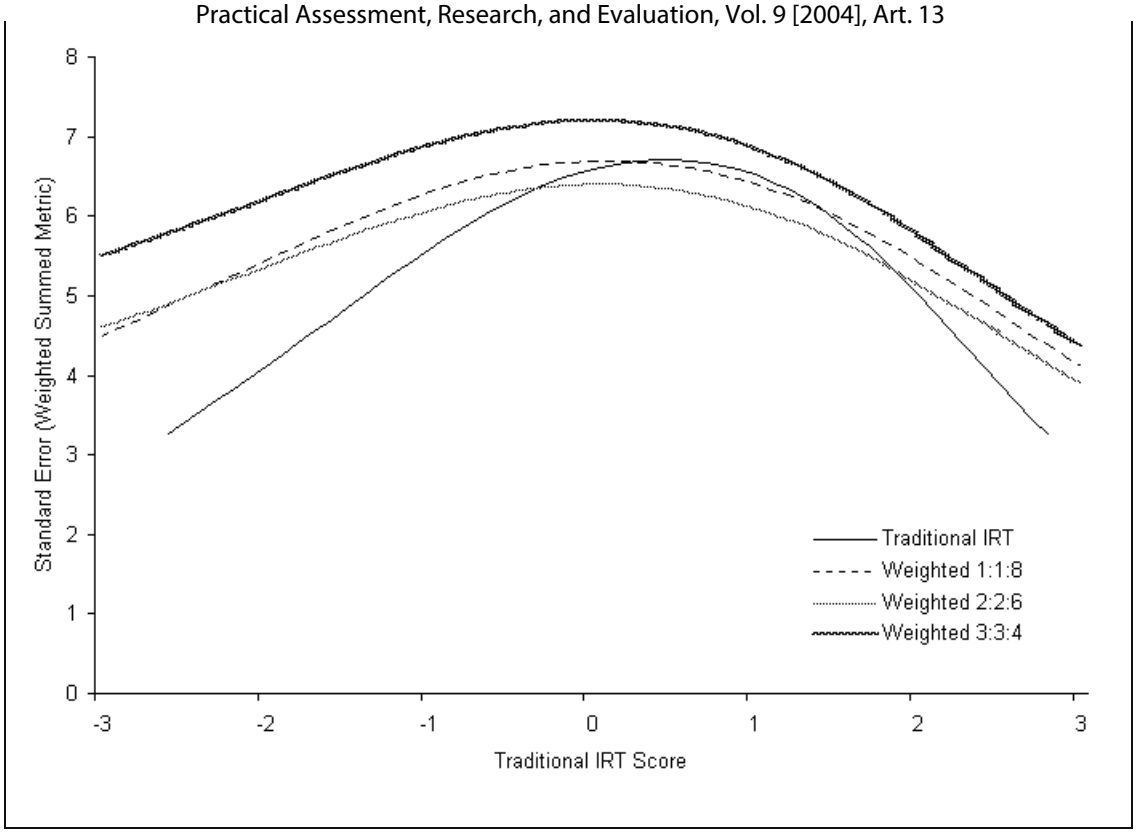


Figure 2. Standard errors for weighted summed scores and IRT scores (Form X).



The standard errors were used to place examinees into a border category if their score estimates plus or minus one standard error included a cutpoint. Based on the IRT scores, 49.4% of the examinees were clearly classified into a performance level (i.e., their score estimate plus or minus one standard error did not include a cutpoint); based on the weighted summed scores using a 2:2:6 ratio, 58.8% were clearly classified. Results were similar for the other weightings.

The patterns of standard errors for Form Y were virtually identical to those for Form X, so the corresponding figures are not included. When one standard deviation above and below each cutpoint were considered in the performance level classifications, 7.2% fewer examinees were clearly classified based on the IRT scoring than based on the weighted summed scoring.

DISCUSSION

Based on these analysis results, what responses can be offered to the four questions posed earlier?

1. What are the correlations between examinees' weighted summed scores and IRT scores?

The 3:3:4 ratio produced the highest correlations with the IRT scores: .933 and .942 for Forms X and Y, respectively. The 2:2:6 ratio produced correlations of .920 and .939 and the 1:1:8 ratio, correlations of .885 and .917. However, the weighted summed scores were even more highly correlated with unweighted summed scores: above .97 for all ratios and forms.

2. If these scores were reported in five categories, what percentage of examinees would be classified differently by the two approaches to computing scores?

Classifications into the performance levels were similar across the two approaches. For Form X and ratio 2:2:6, classifications were the same as those based on IRT scoring for 73.0% of examinees and differed by one level for 25.8%, by two levels for 1.1%, and by three levels for 0.1%. The 3:3:4 ratio produced 75.6% agreement and the 1:1:8 ratio, 67.3% agreement. Results were similar for Form Y.

3. Does one approach favor high or low achievers?

Neither approach consistently favors high or low achievers.

4. Which approach yields a lower standard error of measurement?

The IRT approach yields a slightly smaller standard error at the lower and upper ends of the scale, but the weighted summed score approach using a 2:2:6 ratio yielded slightly lower standard errors in the middle of the scale. For Form X, the 1:1:8 ratio produced standard errors slightly larger than those for the 2:2:6 ratio; the 3:3:4 ratio produced standard errors that were larger still. For Form Y, the 2:2:6 ratio again produced the smallest standard errors, but the 1:1:8 ratio produced the largest.

The results of this study suggest that the effect – on examinee scores, standard errors, or performance level classifications – of explicitly weighting multiple-choice items, short-answer items, and extended constructed-response items to achieve a specified ratio on the raw score metric depends, not surprisingly, on the size of the weights. The ratio of 2:2:6 required weights ranging from 0.88 to 1.29 because the desired ratio was already reflected in the testing time devoted to each type of item and in the numbers of score points for each item and numbers of items. The application of this weighting had little effect on the results. Other ratios required more extreme weights, however, and resulted in larger, though still small, changes in the results. As Table 2 shows, for each form, the ratio that required weights farthest from 1 resulted in the largest standard error; the patterns are less clear for correlations and classification agreement, however.

A few cautions are in order. First, in the scoring stage, different weights were de facto attached to the extended constructed-response items. An item that is scored 1 to 7 has the potential to contribute more to the final score than an item scored 1 to 4. It is possible for all ratings from the extended constructed-response items to have the same weight. This can be accomplished by simply multiplying the weights discussed previously by an item-specific weight.

Another concern is the decision to ignore omitted items in the IRT scoring. A follow-up to this study might treat such items as wrong and compare the results. Given that examinees were not instructed to try to complete the test and, particularly, the low motivation conditions in a field test, it is not clear that omitted items should be seen as evidence of low achievement. Data from an operational administration might also be analyzed based on a different decision on this issue.

Additionally, as Rudner (2001) observed, assessing the effect of weights on a test's reliability is not the same as assessing the effect on validity. In this study, standard errors were computed as an index of score reliability. As no criterion or other data against which to compare the scores were available, it was not possible to examine the effect of the weights on validity.

Finally, the desired weighting is based on ratios on the raw score metric and does not take into account other characteristics of the item types, such as the variance of the scores. Arguments could be made for alternative explicit weighting schemes that would use other criteria for determining the contribution of each item type. The decision to use ratios on the raw score metric was based on a desire to be able to report to the public that each item type did count for the specified percentage of "points" on the scale.

CONCLUSION

The results presented above suggest that the effect of weighting by item type depends on the size of the weights. The ratio that most closely approximated the ratios of testing time and numbers of score points for each item type produced results most comparable to those of IRT scoring. The ratios that required weights farther from 1 resulted in larger standard errors. However, in the score range where most examinees fall, the standard errors are similar across scoring approaches. The effect of changing scoring approaches is quite small for most examinees.

If the intended relative emphasis on different item types is reflected in the test's design, then it is likely that the effort required to impose explicit weights will be disproportionate to their effect. As the difference between the intended emphasis and the test's design increases, however, the effect of the weighting will also increase. In other words, the effect of explicit weights is likely to be greatest when a test is least well-designed for its intended use.

Acknowledgments

We are grateful to the Education Quality and Accountability Office for permission to use the data analyzed in this study. Connie Kidd and Kelsey Saunders, both of the EQAO, provided much-appreciated information and advice.

This article is based in part on a paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA, 4 April 2002.

Correspondence concerning this article should be addressed to Ruth A. Childs, OISE/UT, 252 Bloor Street West, 11th Floor, Toronto, Ontario M5S 1V6. E-mail: rchilds@oise.utoronto.ca

References

Baker, F. B. (1992). *Item response theory: Parameter estimation techniques.* New York: Marcel Dekker.

Ban, J.-C., & Lee, W.-C. (2002, April). *Psychometric properties of composite scores and effects of differential weights using different combining methods.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Childs, R. A., Elgie, S., Gadalla, T., Jaciw, A. P., & Traub, R. E. (2001, June). *Analyses of the EQAO's Grade 9 Assessment of Mathematics field test data. Final report*. Toronto, ON: Authors.

Childs, R. A., Jaciw, A. P., & Saunders, K. (2002, June). *Marking code alignment: Combining markers' judgments with*

item parameter estimates to set performance level cutpoints. Paper presented at the annual meeting of the Psychometric Society, Chapel Hill, NC.

Cohen, J., & Jiang, T. (2001, June). *IRT scores and the standard errors with differentially weighted items.* Unpublished manuscript.

Ercikan, K., Schwarz, R. D., Julian, M. W., Burket, G. R., Weber, M. M., & Link, V. (1998). Calibration and scoring of tests with multiple-choice and constructed-response item types. *Journal of Educational Measurement, 35,* 137-154.

Gulliksen, H. (1950). *Theory of mental tests.* New York: Wiley.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory* (Measurement methods for the social sciences, Vol. 2). Thousand Oaks, CA: Sage.

Ito, K., & Sykes, R. C. (2000, June). *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, Utah.

Kolen, M. J., & Wang, T. (1998, April). *Conditional standard errors of measurement for composite scores using IRT.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

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

McDonald, R. P. (1968). A unified treatment of the weighting problem. *Psychometrika, 33,* 351-381.

Muraki, E., & Bock, R. D. (1996). *PARSCALE: IRT based test scoring and item analysis for graded open-ended exercises and performance tasks* (Version 3.0). Chicago, IL: Scientific Software International.

Rudner, L. M. (2001). Informed test component weighting. *Educational Measurement: Issues and Practice, 20* (1), 16-19.

Samejima, F. (1969). Estimation of ability using a response pattern of graded scores. *Psychometrika Monograph,* No. 17.

Sykes, R. C., Truskosky, D., & White, H. (2001, April). *Determining the representation of constructed-response items in mixed-item format exams.* Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA.

Thissen, D., Nelson, L., Rosa, K., & McLeod, L. D. (2001). Item response theory for items scored in more than two categories. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 141-186). Mahwah, NJ: Erlbaum.

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.

Wang, M. D., & Stanley, J. C. (1970). Differential weighting: A review of methods and empirical studies. *Review of Educational Research, 40,* 663-705.

Wilson, M., & Wang, W. (1995). Complex composites: Issues that arise in combining different modes of assessment. *Applied Psychological Measurement, 19,* 51-71.

Descriptors: Error of Measurement; Reliability; Standard Error; Item Response Theory; Weighting

Citation: Childs, Ruth A., Susan Elgie, Tahany Gadalla, Ross Traub, Andrew P. Jaciw (2004). Irt-linked standard errors of weighted composites. *Practical Assessment, Research & Evaluation*, 9(13). Available online: http://PAREonline.net/getvn.asp?v=9&n=13.