ITEM RESPONSE THEORY FOR WEIGHTED SUMMED SCORES

Brian Dale Stucky

A thesis submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Master of Arts in the Department of Psychology.

Chapel Hill
2009

Approved by:
David Thissen, Ph.D.
Robert C. MacCallum, Ph.D.
A.T. Panter, Ph.D.

ABSTRACT

BRIAN STUCKY: Item Response Theory for Weighted Summed Scores
(Under the direction of David Thissen)

Tests composed of multiple sections are routinely weighted by section. Methods for weighting have existed perhaps as long as there have been tests; however, computing and evaluating the quality of weights has not evolved with recent advances in test theory (Item Response Theory (IRT)). While IRT may be used to compute accurate estimates of ability based on a variety of information (e.g., pattern responses or summed scores), there has been little research on the computation of scale score estimates for tests with arbitrary item or test section weights. The present work provides an extension to a recursive algorithm for the computation of IRT-scale scores from weighted summed scores.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

CHAPTER 1

A HISTORY OF ITEM WEIGHTING

Item weighting has historically received much attention. In tests constructed and evaluated using classical test theory (CTT), items have been given differential weights, allowing them to contribute more or less to a composite measure. Methods for selecting these weights have varied considerably; early methods attempted to account for items with differing length, difficulty, or assumed validity, etc., while more modern conventions select weights based on the statistical properties of the items.

Item weighting has rarely been considered in an item response theory (IRT) framework (see Sykes & Hou, 2003 for an exception). For IRT scale scores, the item parameter estimates serve as implicit weights, where guessing, item difficulty, and item discrimination parameters are used to provide the best estimates of the ability of individuals taking a test. Recent test theory research has focused on IRT models, with implicit item weighting; the literature on explicit weighting has become an historical aside, in which references to item weighting are made within the CTT domain.

The present paper seeks to understand how IRT ability estimation might be conducted for tests with a priori item weights. This would be implemented in situations where a testing program has explicitly chosen weights that are applied prior to IRT analysis. The methods contained in this paper would allow weighted summed scores to have corresponding IRT scale scores. Consequences of this procedure will be discussed in relation to how item

weights affect the resulting estimates of individuals' ability and the corresponding levels of error (or uncertainty) across the ability continuum. To better relate item weighting with IRT, a brief review of traditional weighting techniques is provided as a precursor to the methods for combining the two domains (weighting and IRT).

Weighting in Classical Test Theory

Item weighting has a long history in CTT. Broadly speaking, in situations in which items have obviously heterogeneous characteristics, test administrators have routinely accounted for these differences through item weighting. Weights have often been applied with respect to item difficulty, length, number of possible responses, etc. As will be demonstrated later, weights can also be applied to entire test sections. Large scale tests composed of multiple test sections are common (e.g., the Advanced Placement (AP) exams and the General Education Development (GED) tests). For tests with multiple sections, administrators commonly incorporate weights to account for, or to correct for, perceived inequalities between test sections. Many have noted, for example, that constructed response (CR) questions are costly in many ways: (1) they are time consuming, (2) they can be affected by abilities other than the intended dimension, and (3) they offer poor reliability as compared to multiple choice (MC) items (Wainer & Thissen, 1993; Lukhele, Thissen, & Thissen, 1994; Thissen, Wainer, & Wang, 1994).

The choice to weight one item or section more than another usually results from a belief that the item or section with the larger weight is "better," or best reflects the underlying latent trait, and thus should contribute more to a test taker's final score. Typically, however, weights are chosen for surface-level characteristics (e.g., item length, difficulty, and assumed validity), though test analysts may prefer to weight based on the psychometric properties of

the items or test sections. In these cases, weighting based on the test's section and composite reliability and validity is preferred. Deciding how much weight an item or section receives can be determined using either "arbitrary" or "statistical" strategies, or both.

There are many arbitrary methods of item weighting. Researchers have weighted based on item length, the time allowed to complete items, and a "proportional method" based on restricting a test's separate sections to contribute equally to the composite score. As will be clarified, all of these techniques are derivatives of a single general principle. Regarding weighting by length, a commonly seen example is for tests with mixed item types to weight essay questions more heavily than other non-essay questions. This is done not only because essays are implicitly considered by some to contain more valid information about the test taker, but also because a single essay takes far longer to answer than does a multiple choice question. Another approach to weighting tests with different item types is to weight in order to achieve equal contributions to the score for each of several test sections. For example, a 10 item multiple choice test section scored correct/incorrect (10 points possible), is coupled with a single essay question (5 points maximum). To achieve equality a weight of 2 is attached to the essay question (MC_{max}/CR_{max}). The commonality between weighting by length, time allowed, or equality of points is that tests with essay sections are often scored using fewer points than are ultimately allocated after weighting. This convention is practical given that essays are typically judged by multiple raters who must achieve a level of agreement, which is more reasonable to maintain with reduced scoring ranges. Thus, it is often necessary to weight CR items in order to ensure that they contribute adequately to a test.

While arbitrary weighting methods are still popular, other statistical methods may provide more useful weights for test sections and items. One such method is weighting based on a multiple correlation approach; this approach has been among the most widely used methods. The limiting factor of this technique is the existence of a criterion variable (x_0) that is predicted from the other items in the test. If such a criterion variable exists, the β_i-weights form the best possible linear combination of items ($x_1, x_2 \ldots x_n$) for predicting $\hat{x}_0 = \beta_{01.23\ldots n} x_1 + \beta_{02.13\ldots n} x_2 + \beta_{03.12\ldots n} x_3$. The weights, β_n's, are largest for items that are strongly associated with the criterion item and also for items that are independent from other items (Wilks, 1938). As a measure of effect size, R^2 is the maximized quantity of multiple regression weights (McDonald, 1968). However, problems with generalizing multiple correlation models with many parameters and small sample sizes are relevant in this context (Cohen, Cohen, West, & Aiken, 2002). "Shrinkage" is also particularly relevant in these situations because R^2 can be over-estimated by capitalizing on the properties of the sample (Cohen, Cohen, West, & Aiken, 2002).

A different approach to weighting, "reliability weighting," has also been considered. It is clear that for tests composed of more than two sections there can be marked differences in the reliabilities of the separate sections. "Reliability weighting" is concerned with finding the optimal weights that are associated with a composite score's maximum reliability, and involve increasing the contribution (weight) of the section with the highest reliability. The procedures for finding the maximum reliability involve matrix manipulations of the test sections' reliabilities and the intercorrelation between the sections (see Peel, 1948 for the matrix derivations and Wainer & Thissen, 2001, chap. 2 for a practical application).

Weighting in Item Response Theory

Arbitrary weighting methods clearly affect the reliability of measures from a CTT perspective. Of present interest is how arbitrarily weighting items and test sections might affect a test with IRT analysis. Lukhele and Sireci (1995) discussed this problem in the context of the conversion of the writing skills section of the GED test from CTT analysis to an IRT analysis. Traditionally, the test had weights of .64 and .36 for the MC and CR sections, respectively, which were arbitrarily chosen to allow the essay section to adequately contribute without overly reducing the composite reliability (like most tests, we can assume the reliability of the CR section to be substantially less than that of the MC section). Lukhele and Sireci obtained the "unweighted" IRT marginal reliabilities (Green, Bock, Humphreys, Linn, & Reckase, 1984; Thissen & Orlando, 2001, p. 119) for each section, followed by a "maximum" reliability, in the IRT sense of the word, for the composite as computed by a simultaneous IRT analysis of all items. From these IRT marginal reliabilities, the traditional .64/.36 weights were applied to the trait estimates to compute a "weighted" composite reliability, which was nearly 6 percent less than the unweighted IRT "maximum" reliability. This loss in reliability can be in part attributed to weighting the test section marginal reliabilities by .64/.36, when "optimal" IRT weights suggest a split of .86/.14, indicating that the MC section should be worth more than 6 times as much as the CR section.

While this procedure is an important early step in investigating weighting and IRT, Lukhele and Sireci's method of combining IRT scores with weights ignores a central component in IRT. Reliability is rarely used in the context of IRT because, unlike CTT, IRT allows for individual items to have varying implicit weights which change as a function of the item's IRT parameters. Lukhele and Sireci modified the differential weights of each item

when using the marginal reliabilities of the test sections to produce an "IRT weighted" composite reliability. It should then come as no surprise that applying weights in an a posteriori fashion to the estimates of ability reduced overall reliability.

Sykes and Hou (2003) used a more direct approach to combine weighting with IRT. Sykes and Hou demonstrate that for tests composed of combined item types (CR and MC), weights may be applied prior to IRT estimation of scores. This weighting was accomplished by increasing the portion of the test characteristic curve (TCC) that was contributed by CR items and then using the modified TCC to create a weighted-summed-score to IRT-score conversion table. In the example considered by Sykes and Hou, all CR items were weighted by 2, with the MC items receiving unit weights.

Sykes and Hou considered the effectiveness of this strategy by comparing differential standard errors across the ability continuum. In this case, when compared to the response pattern estimates, slight increases in standard errors were seen across average and upper levels of ability, with the exception occurring for lower scores, where weights reduced estimation error. A potential explanation for the effect Sykes and Hou report is that the CR items, as compared to the MC items, happen to be somewhat easier and probably discriminated better among individuals at lower ability levels, though this must be speculation because item parameters are not reported.

In the example used by Sykes and Hou there is a clear intent to increase the contribution of CR items. It is often argued that the perceived benefit of increased test composite validity for positively weighted CR sections overshadows the cost associated with a decrease in test reliability (Kane & Case, 2004; Rudner, 2001). It is true that if such discrepancies in validity exist, then weighting based on validity would seem a reasonable approach (Lord & Novick,

1968). However, little empirical evidence exists showing such a disparity, while overwhelming evidence supports clear differences in reliabilities for CR and MC test sections. While weighting to satisfy either condition seems paradoxical, Wainer and Thissen (2001), demonstrate that in CTT, when considering the optimal reliability of test, for every situation of more than one test section, there is a range of potential weights that vary between those that produce the optimal reliability and those the produce a reliability no less than the most reliable test section. Thus, even in situations where a CR test section has low reliability there is still a range of possible weights which provide (at least) as good reliability as if the section were removed.

Often with decisions regarding weights there are no clear answers. It is not uncommon for weights to produce slight increases in performance of one aspect of a test at the expense of the performance of another area. For example, the weights Sykes and Hou chose increased the contribution of the CR section by 16%, resulting in improved test information at the lower score ranges, but reduced the marginal reliability of the composite by 2%. Many tests contain weighted CR sections, and given the small amount of literature on the subject (e.g., two known papers present IRT approaches) and the lack of clear guidelines for the choice of weights, the current investigation of how IRT ability estimation may be conducted directly from weighted summed scores seems timely.

Weighted Summed Scores for Mixed Item Types

To appropriately incorporate item weighting with IRT, new methods are needed. Specifically, while Lukhele and Sireci applied arbitrary weights directly to the expected a posteriori estimates of ability (EAPs), and Sykes and Hou weighted the observed item responses in the TCC, a true incorporation of item weights would allow for separate IRT ability estimation to occur for each weighted summed score. This would permit weighted summed scores to have associated EAPs and corresponding standard deviations (SDs). While little is known about the relation of weighted summed scores with IRT ability estimation, one can begin by considering weighted summed scores as an extension of simple summed scores for tests with multiple item types. For the case in which a test has only a single section, the number of maximum possible summed scores is:

$$n_i * (c_i - 1) + 1. \quad (1)$$

Here n refers to the number of items i in a test section with c number of response options per item. For example, a 4 item test composed of binary items has 5 possible summed scores, (0,…, 4). For tests with s sub-sections, the maximum possible summed score is $\sum_{1}^{s} n_s (c_s - 1)$. Thus, if the 4-item binary test is combined with a CR section with 2 items worth 2 points each, the maximum possible summed score would be 8. Often tables of summed scores are helpful. Here, Table 1 illustrates that the total number of possible summed scores for this test is 9 (0, … , 8).

Table 1. Possible summed scores of a test composed of two CR items with three response categories and four MC items

	CR max. possible summed scores					
MC max. possible summed scores	Sum	0	1	2	3	4
	0	0	1	2	3	4
	1	1	2	3	4	5
	2	2	3	4	5	6
	3	3	4	5	6	7
	4	4	5	6	7	8

For all scores other than the minimum (0) and the maximum (8), individuals can obtain a given summed score with more than one pattern of MC and CR summed scores. For example to obtain a 1 a test taker must incorrectly answer all items except one binary item or receive a 1 on a CR item and 0's elsewhere. As seen in Table 1, for unit-weighted test sections there are at least two ways of obtaining a given summed score for all possible scoring responses other than the minimum and maximum summed scores.

However, for tests composed of multiple weighted sections, the maximum number of unique weighted summed scores is less clear. The process for determining the number of summed scores can again be imagined from the standpoint of a similar table where item responses now have weights. Given the same test example previously used, when weights of 1.2 are chosen for the MC section the range of possible weighted summed scores are (0, 1.2, 2.4, 3.6, 4.8); when weights of .8 are used for the CR section the possible weighted summed scores are (0, .8, 1.6, 2.4, 3.2). Note that the number of summed scores for the individual sections is unchanged from the unit-weighted example. However, as illustrated in Table 2, combining these weighted summed scores to obtain the total number of possible weighted summed scores for the test produces more unique summed scores than were previously possible.

Table 2. Possible weighted summed scores of a test composed of two CR items with three response categories and four MC items

MC max. possible weighted summed scores	CR max. possible weighted summed scores				
Sum	0	.8	1.6	2.4	3.2
0	0	.8	1.6	2.4	3.2
1.2	1.2	2.0	2.8	3.6	4.4
2.4	2.4	3.2	4.0	4.8	5.6
3.6	3.6	4.4	5.2	6.0	6.8
4.8	4.8	5.6	6.4	7.2	8.0

In this case there are 13 possible unique weighted summed scores. This occurs because 6 different weighted summed scores can be obtained from 2 separate combinations of CR and MC items (which are illustrated in grayscale). If, the weights chosen are "sufficiently unique" (i.e., numbers that avoid proportionality), then the possible weighted summed scores is the number of rows (the number weighted CR summed scores) multiplied by the number of columns (the number of weighted MC summed scores), or for *s* test sections:

$$\prod_{1}^{s}(n_s(c_s-1)+1) \tag{2}$$

which in this case is (4*(2-1)+1)*(2*(3-1)+1) = 25. Presenting weighted summed scores for tests with multiple weighted test sections takes the form of a cube for tests with 3 sections, or a hyper-cube for tests with more than 3 sections.

More generally, the problem of determining the number of unique weighted summed scores is actually a problem of determining how many integral multiples correspond between weighted test sections. While computing the scores of the composite, any multiples of test section weights that may be derived from more than one combination of weights will result in a duplicate, or redundant weighted summed score. The total number of weighted summed scores is then found by counting every weighted summed score once (i.e., counting each unique score only once). It should be noted that in cases other than the simplest, the process

of determining the number of weighted summed scores is only reasonably done with a computer. For example, a mixed item-type test with 98 weighted binary items and 4 weighted CR items (11 scoring categories) produced 4059 weighted summed scores, which required nearly 6 minutes to compute on a 2.13 GHz Duo Core processor using the interpreted statistical language R.

To conduct IRT scoring, the likelihoods associated with these weighted summed scores require computation. The challenge is that for tests with many items and weights, obtaining these likelihoods can be computationally difficult. This paper presents a method for obtaining IRT scores from weighted summed scores and describes some properties of scoring tests with weights when using item response theory.

CHAPTER 2

AN EXTENSION OF THE RECURSIVE ALGORITHM

While IRT scale scores are most often associated with each pattern of item responses, and such scores are optimal if the model describes the data, in many contexts users of test results are more comfortable with scores that are based on sums or weighted sums of values associated with each response. It is well-known that IRT can be used to compute scale scores associated with each simple summed score (Thissen, Pommerich, Billeaud, and Williams, 1995; Thissen & Orlando, 2001; Thissen, Nelson, Rosa, & McLeod, 2001); in the process, one can compute not only the scale score but also an estimate of its standard error, and the modeled proportion expected to obtain each summed score, which can be used to compute modeled percentiles and check the goodness of fit of the observed distribution of summed scores. These values are all computed using previously-estimated item parameters; the challenge in their computation is to compute the likelihood over θ for each summed score. An extension of the methods described in this chapter provides an algorithm to compute the IRT likelihood for weighted summed scores.

Thissen, Pommerich, Billeaud, and Williams (1995) illustrate the steps needed to compute the likelihood of each summed score given θ. The likelihood of any summed score x (the sum of item scores $u = 0,\ldots,U_i$ for all items i):

$$L_x = \sum_{\ni x=\sum u_i}^{\text{response patterns}} L(\mathrm{u} \mid \theta) . \qquad (3)$$

Considering only the items, the likelihood for every response pattern is

$$L(u \mid \theta) = \prod_i T_{u_i}(\theta) \tag{4}$$

in which $T_{u_i}(\theta)$ represents the conditional probability of responding in category u_i on item *i*. For the two-parameter logistic (2PL) model, $T(u_i = 1 \mid \theta)$, for a correct response, takes the form

$$T(u_i = 1 \mid \theta) = \frac{1}{1 + \exp[-a_i(\theta - b_i)]}, \tag{5}$$

while an incorrect response has the probability: $T(0 \mid \theta) = 1 - T(1 \mid \theta)$, or $T_0(\theta) = 1 - T_1(\theta)$ in more compact notation. The probability of a correct response 1 to item *i* given the level of the latent variable, θ, is a function of the "discrimination" power of the item (a_i) and the difficulty of the item (b_i).

For items composed of *m* ordered response categories, Samejima's (1969) graded response model (GRM) describes the probability of a responses in category *k* or higher, where $k = 0, 1, 2 \ldots, m - 1$:

$$T^*(k \mid \theta) = \frac{1}{1 + \exp[-a_i(\theta - b_{ik})]}, \tag{6}$$

noting that $T^*(0 \mid \theta) = 1$ and $T^*(m \mid \theta) = 0$. Here the probability of responding in category *k* is the difference between the probabilities of responding in *k* or higher and the higher response: $T_i(k \mid \theta) = T_i^*(k \mid \theta) - T_i^*(k+1 \mid \theta)$.

Combining equations 3 and 4 provides the likelihood for every summed score *x*:

$$L_x(\theta) = \sum_{\ni x = \sum u_i}^{\text{response patterns}} \prod_i T_{u_i}(\theta) \tag{7}$$

Taking the product of the likelihood of score *x* and the population density φ(θ) provides the posterior distribution of score *x*, which has an associated probability computed by

$$P_x = \int \sum_{\ni x = \sum u_i}^{\text{response patterns}} \prod_i T_{u_i}(\theta)\phi(\theta)d(\theta)\,. \qquad (8)$$

Finally, numerical integration using the posterior distribution and the probability of score x is used to compute the expected a posteriori (EAP), or the mean of the posterior distribution of each summed score, along with the standard deviation (SD) corresponding to each EAP.

Lord and Wingersky, 1984 described a recursive algorithm that can be used to compute the likelihood of summed scores given θ. The recursive algorithm is easiest to describe with dichotomous items: the probability that an individual correctly answers a test with a single dichotomous item (u = 1) is T_1 where the complement represents the probability of an incorrect answer (1 - T_1), so that the summed score likelihood for this test, L_x^1, is $T_{x_1}(\theta)$.. When a second item is added to the test there are three possible scores (x = 0,…,2). The likelihood of x = 0 is the product of the previous item complement with the current item complement, (1-T_1)*(1-T_2); the likelihood of x = 2 is T_1T_2, and finally, the likelihood of x = 1 is $T_1(1\text{-}T_2)$ + $(1\text{-}T_1)T_2$. The likelihood of a summed score that is not 0 or 2 for this two item test is found by adding the product of the likelihood of the score for the preceding item ($L_x^{I^*}$) and responding incorrectly to the present item ($T_{0_{I^*+1}}$) to the product of the likelihood and of one less than the summed score ($L_{x-1}^{I^*}$) and responding correctly to the present item ($T_{1_{I^*+1}}$).

The recursion in this algorithm can be viewed as an updating process in which the likelihoods of all possible summed scores as a function of θ are reevaluated for every succeeding item. Thissen et al. (1995) present this algorithm for any number of response categories, so that for all items (0… I*), the summed score is $x = \sum u_i$ and the likelihood of

summed score x is $L_x^{I^*}(\theta)$. The recursive algorithm begins by setting the first item equal to zero ($I^* = 0$). For the first item the likelihoods associated with each score x is:

$$L_x^{I^*} = T_{x_{I^*}}(\theta). \tag{9}$$

The likelihood for the next item, $I^*+ 1$, is computed by:

$$L_{x+u}^{I^*+1} = \sum_{u_{I^*+1}} L_x^{I^*}(\theta) T_{u_{I^*+1}}(\theta), \tag{10}$$

which is repeated until I^* is I.

IRT for Weighted Summed Scores

In order to incorporate weights into IRT score estimates, an extended version of this recursive algorithm may be used. In the original recursive algorithm, at each successive iteration, the maximum possible summed score increases exactly one point for binary items, or one less than the number of categories for graded items. However, if two test sections receive differing weights, the increases in summed scores may not be integers; the set of summed scores for these types of tests could have many values. For unit-weighted tests there are many response patterns with the same summed scores but for tests with differential weights, far fewer response patterns may be associated with each weighted summed score, and the number of potential response patterns can be different for each summed score. There are often many more possible weighted summed scores than there are unit-weighted summed scores, and the distances between the weighted summed scores are not uniform, as they are with summed scores.

The recursive algorithm to compute scale scores for weighted summed scores may be thought of as a two-stage process that is a re-expression of the original algorithm used for unit weighted summed scores. In stage 1, as in equation 10, likelihoods are computed for

each weighted summed score which are the sum of the products for the trace lines of each pattern of scores that forms the weighted summed score. In the general weighted case the patterns of scores are not collections of integers, but rather are "item scores" (which may be section weights). When this stage is completed a test with separately weighted sections has likelihoods associated with each possible weighted summed score. In stage 2 the likelihoods are "collapsed," or, summed together for each combination that has the same weighed summed score. If a weighted summed score can only be computed by one combination of test section scores, then the likelihood for that combination of scores remains as it was.

Many factors affect the number of weighted summed scores, including the uniqueness of the weights and the number of weighted sections. Often the number of weighted summed scores is very large. Recall that a mixed item-type test with 98 weighted binary items and 4 weighted CR items (11 scoring categories) produced 4059 weighted summed scores. In these situations, where score reporting may be difficult, an additional procedure may be implemented to transform the weighted summed scores back into integer scores. In this optional, third step, the likelihoods are further collapsed for weighted summed scores that round to a given integer, resulting in as many integer scores as there are unit-weighted summed scores if the weights are normalized to sum to the total unit-weighted summed score.

As an example of these procedures, consider the test described previously: the first MC section includes 4 items with weights of 1.2, and the second CR section includes 2 items with three scoring categories and weights of .8. The extended recursive algorithm is used to compute the likelihoods for all 25 possible combinations of item scores (i.e, 0, 0.8, 1.2, 1.6, 2.0, ... ,8.0). For the weighted summed scores 0.0 and 8.0 the EAPs are the same as would be

computed from the original algorithm, because the likelihoods are the products of the same trace lines (namely, all incorrect or all correct). However, the extended recursive algorithm distinguishes between the weighted summed scores 0.8 and 1.2 (which are the same for unit-weighted summed scores). The likelihoods for score 0.8 are computed only from those patterns of scores in which all the MC items were incorrect and only the first scoring category is achieved for one CR item, and vice versa for the score 1.2. Stage 2 of the algorithm considers the full list of weighted summed scores and combines the likelihoods for scores that are computed from more than one combination of item scores. For example, a score of 2.4 may occur either by correctly answering two MC items, and scoring a zero elsewhere, or by receiving a score of 2.4 on the CR section, and scoring a zero elsewhere.

An Example of Using IRT to Score Weighted Test Sections

Using the example above, weighted IRT scale scores were computed from four MC items and two CR items with three response categories per item. Because scoring is conducted after parameter estimation, we suppose here that a 1PL model (a 2PL model with equal slope parameters) was fitted to the MC items and the GRM to the CR section (see Table 3 for the IRT parameters).

Table 3. IRT parameters for test sections with 1PL and GRM items

Item	a	b_1	b_2
MC_1	2.5	-1.00	
MC_2	2.5	-0.33	
MC_3	2.5	0.33	
MC_4	2.5	1.00	
CR_1	1.5	-1.50	-0.75
CR_2	1.5	0.75	1.50

A priori weights of 1.2 were used for the MC section and 0.8 for the CR section. Scoring was conducted using the original algorithm for unit-weighted summed scores and the extended recursive algorithm for weighted summed scores. For both unit and arbitrary weights, Table 4 contains EAPs for each combination of item scores by test section.

Table 4. Summed scores, weighted summed scores,

and EAPs for all score combinations of 1PL and GRM items

	CR summed scores					
	Sum	0	1	2	3	4
MC summed scores	0	0	1	2	3	4
	1	1	2	3	4	5
	2	2	3	4	5	6
	3	3	4	5	6	7
	4	4	5	6	7	8
	CR weighted summed scores					
	Sum	0	0.8	1.6	2.4	3.2
MC weighted summed scores	0.0	0.0	0.8	1.6	2.4	3.2
	1.2	1.2	2.0	2.8	3.6	4.4
	2.4	2.4	3.2	4.0	4.8	5.6
	3.6	3.6	4.4	5.2	6.0	6.8
	4.8	4.8	5.6	6.4	7.2	8.0
	EAPs for MC and CR score combinations					
EAPs for MC and CR score combinations		-1.62	-1.28	-1.06	-0.76	-0.68
		-0.90	-0.72	-0.49	-0.25	-0.16
		-0.35	-0.23	0.00	0.23	0.35
		0.16	0.25	0.49	0.72	0.90
		0.68	0.76	1.06	1.28	1.62

Note: the top panel contains unit-weighted summed scores for MC and CR test sections, the middle panel contains the weighted summed scores for MC weights of 1.2 and CR weights of 0.8, the bottom panel contains the EAPs for each MC and CR score combination regardless of item weight.

The EAPs in lower panel of Table 4 represent the likelihoods that form the basis of later score combinations. For example, in the unit-weight case, the EAP for a summed score of 1.0 is computed by combining the likelihoods for MC and CR scores [1,0] and [0,1] (Table 4, top panel). However, in the weighted-summed score case, the EAPs for 0.8 and 1.2 remain separate scores (Table 4, middle panel). Computing unit-weighted summed score EAPs involves combining the likelihoods associated with the top panel of Table 4 for all cases in which a given element forms a summed score x. As the middle panel of Table 4 makes clear, depending on the weights, the weighted summed scores might involve unintuitive score combinations. For instance, an EAP for the weighted summed score 3.6 would involve combining the likelihoods for item scores [3,0] and [1,3]. The effect of weighting on score estimation will be discussed in the following sections.

In the present example, the extended recursive algorithm was used to compute EAPs and SDs separately for both the unit and arbitrarily weighted cases (Figure 1).

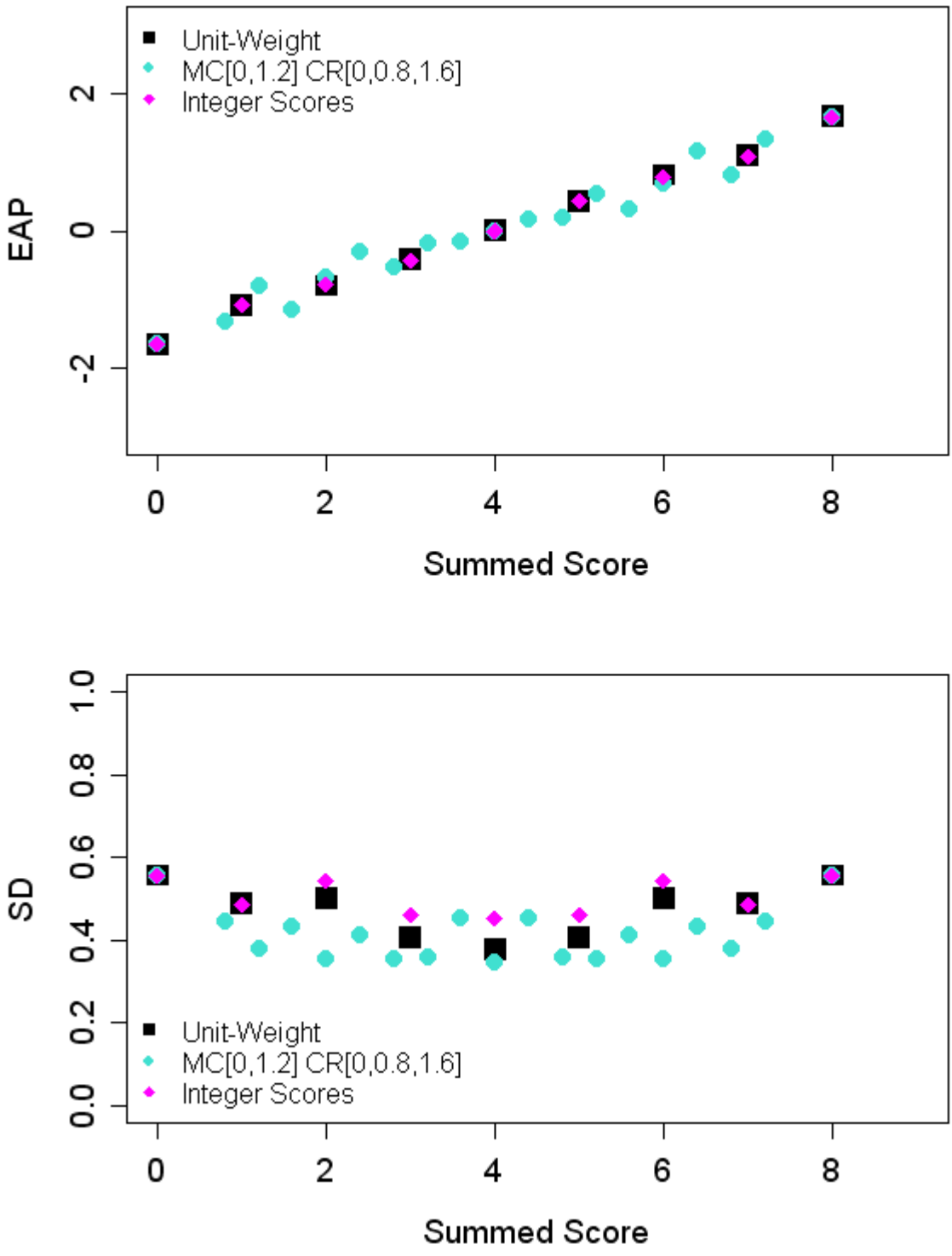

Figure 1. EAPs and SDs of IRT scale scores for unit and weighted summed scores

Note: this figure contains the EAPs (top panel) and corresponding SDs (bottom panel) for unit-weighted summed scores, weighted summed scored, and integer scores.

The lower panel of Figure 1 indicates the EAPs for these weighted summed scores, as compared to unit-weighted summed scores, have less uncertainty for the majority of the range of scores. Explanations for this will be proposed in later sections. Also notable, from the top panel in Figure 1 is that while EAPs for both the unit and arbitrarily weighted summed score are closely related, the EAPs for the weighted summed scores are not monotonically increasing. In other words, increases in weighted summed scores do not necessarily indicate increases in EAPs. The EAP associated with a weighted summed score of 1.2 is greater than that for 1.6. This reflects the relative importance of the MC items in determining levels of ability (see Thissen, Nelson, & Swygert, 2001, pp. 303 - 307 for a discussion of relative weights). Finally, the upper panel of Figure 1 indicates that collapsing weighted summed scores into integer scores provides EAPs which are nearly identical to unit weighted summed scores. However, with the exception of scores 0, 1, 7, and 8, which are produced from the same likelihoods for both integer and summed scores, the SDs of the EAPs are slightly higher for the integer scores. Referring back to Table 4, this may be because the integer scores are combinations of likelihoods which have similar means, but dissimilar distributions.

CHAPTER 3

THE EFFECTS OF WEIGHTING ON SCORING TESTS

In many situations, test developers select weights prior to administering tests and reporting scores. When the weights are known the procedures described in the previous chapter may be used to obtain EAPs and SDs for the weighted summed scores and integer scores. However, an alternative use of the extended recursive algorithm is to compute EAPs and SDs for a variety of weight combinations, which may then be used to evaluate the efficacy of different weighting schemes. To enable such comparisons the recursive algorithm was implemented in the statistical language R to iterate over a range of weights to compare the effects of a variety of hypothetical weight combinations on IRT scores and score distributions.

For tests with a large number of items, or when comparing many weight combinations, graphically evaluating the effects of weights on the EAPs and SDs can be tedious. Thus, the average error variance across scores is computed to reflect the performance of a weighting scheme:

$$\bar{\sigma}^2 = \sum \sigma_x^2 p_x. \tag{11}$$

The average error variance is the sum of the product of the score variance and the probability of a given score. As an overall indicator of the quality of the weights, one minus the average variance is the marginal reliability (Green, Bock, Humphreys Linn, & Reckase, 1984) for standardized θ. When the average variance is low, or when reliability is high, the weights combine likelihoods which result in scores with low error SDs and thus may be considered

better than other sets of weights which result in higher average error variances. When considering multiple weight combinations, the set of weights that produces maximally reliable scores are the "optimal" weights.

The procedures for obtaining optimal weights were implemented for integer scores, because of limitations involved with comparing average error variances across weighted summed scores. For weighting schemes that produce unique weighted summed scores (i.e., scores that do not collapse), the average error variances are the same for all such sets of weights. This occurs because the likelihoods associated with each weighted summed score are unique for all weight combinations produced from unique weights. However, for a range of unique weights, if integer scores are computed, then likelihoods collapse when the corresponding weighted summed scores round to the nearest integer, producing different marginal reliabilities for each weight combination.

Even so, not all changes in weights necessarily produce unique reliabilities. For relatively short tests, small changes in weights may not result in different combinations of likelihoods for the integer scores. When more than one set of weights produce identical combinations of weighted summed scores that round to a given integer, the algorithm combines the same likelihoods for each weight combination and the average score variance remains unchanged. The degree to which the collapsed likelihoods are different across weights corresponds to the change in the average error variance of the scale scores across different weight combinations.

To illustrate the process of selecting optimal weights for integer scores, some examples of iterating the extended recursive algorithm for a variety of IRT models are provided in the examples that follow.

Example 1

The first example of iterating the recursive algorithm over a range of weights uses a 1PL model for four MC items and the GRM for two CR items (with response categories 0, 1, and 2) (see Table 5 for IRT parameters).

Table 5. Example 1: IRT parameters for test sections with 1PL and GRM items

Item	a	b_1	b_2
MC_1	2.0	-0.50	
MC_2	2.0	-0.25	
MC_3	2.0	0.25	
MC_4	2.0	0.50	
CR_1	2.5	-0.75	0.00
CR_2	2.5	0.00	0.75

To select the optimal set of weights, the recursive algorithm was iterated over 21 sets of weights beginning with weights 0.0 and 1.0 for the MC and CR sections, respectively, and increasing the weight of the MC section by 0.05 units until 1.0 was reached (i.e., [0.0, 1.0], [0.05, 0.95],…, [1.0, 0.0]). In this manner, the effect of the weights may be compared at a variety of levels, initially considering only the CR items (MC weight of 0.0) and last considering only the MC items (CR weight of 0.0).

The average reliabilities obtained from these 21 sets of weights are shown in Figure 2. Interestingly, the reliabilities are approximately symmetric around the unit-weighted case. For example, the reliability of including only the CR items (0.66) is nearly identical to the reliability of including only MC items (0.67). There are three sets of weights that produce optimal reliability ([0.45, 0.55], [0.50, 0.50], and [0.55, 0.45] for the MC and CR sections, respectively). Figure 2 illustrates that there is a range of weights associated the optimal, or near-optimal, reliability for the 1PL and GRM test sections.

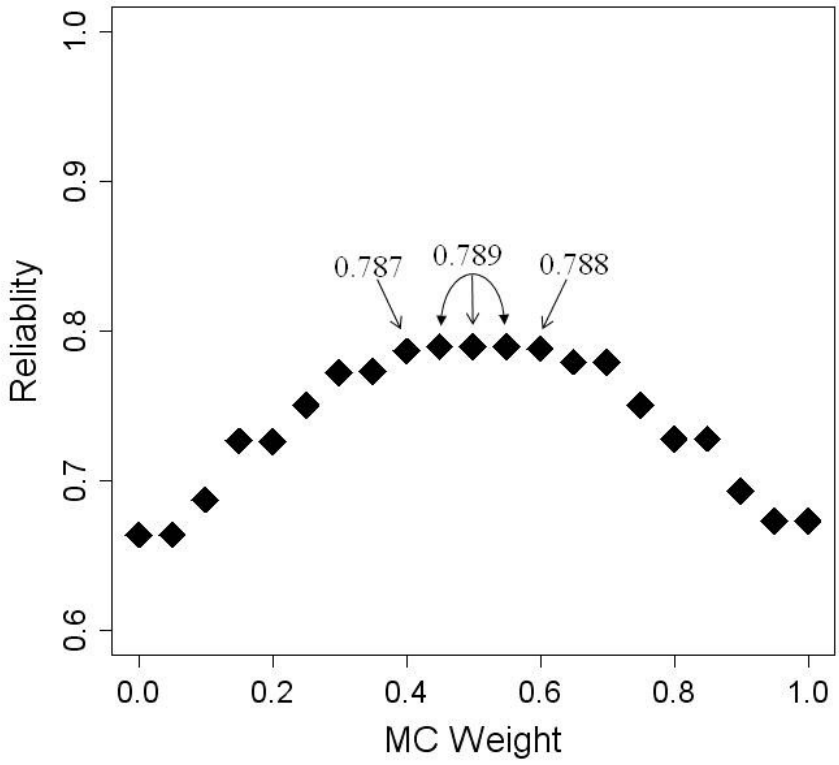

Figure 2. Example 1: Reliabilities for test sections with

1PL and GRM items

As a possible explanation for the symmetry in reliabilities, test information was considered for each section. Test information reflects how precisely the items measure ability over θ. Where information is high, or the standard errors low, the test produces more precise estimates of the examinee's ability. For the 2PL model, test information is

$$\sum_{1}^{n} a_i^2 T_i(\theta)[1 - T_i(\theta)], \tag{12}$$

which is the sum of the item information over *n* items. Item information is the product of the trace lines for a correct and incorrect response and the item's squared slope. The height of information is a function of the discrimination parameter while the location of information is determined by the threshold. Notably, for 1PL and 2PL models, maximum information will always occur where $b_i = \theta$. For binary logistic models, item information is unimodal. For polytomous IRT models, information is a function of the spacing between thresholds, and when thresholds are spaced moderately far apart, there may be as many modes as there are thresholds.

Returning to Example 1, test information was plotted for both sections (Figure 3).

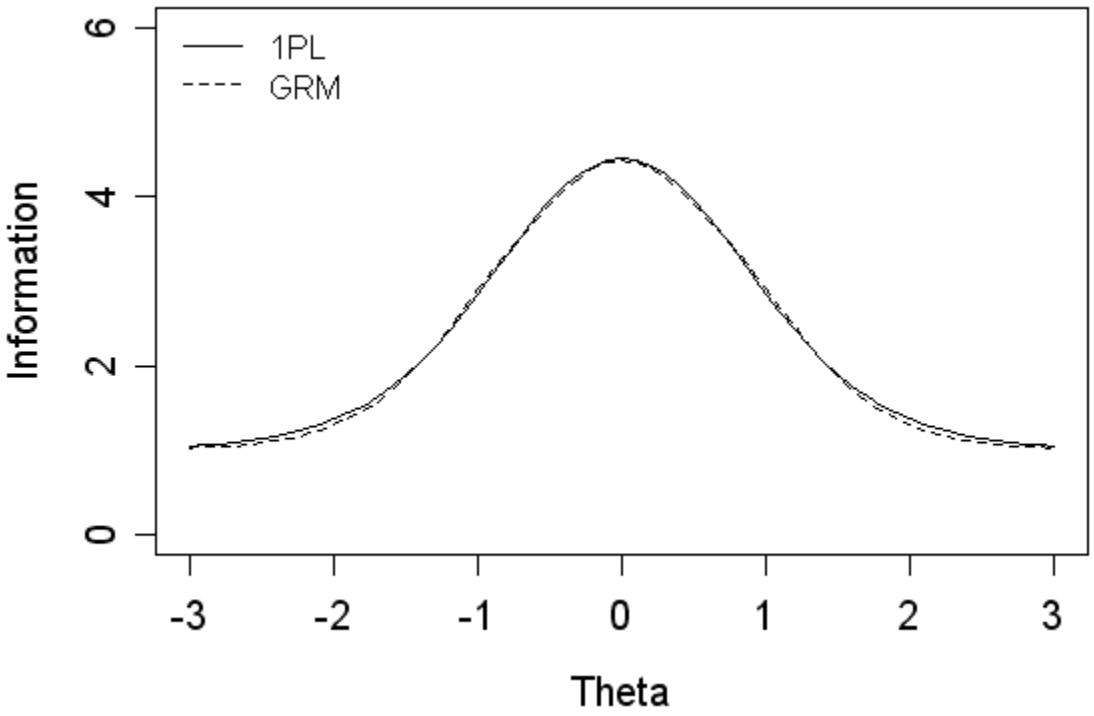

Figure 3. Example 1: Test information for 1PL and GRM items

The item parameters for Example 1 were chosen in such a way that the information functions for the 1PL and GRM sections were nearly identical. The symmetry seen in Figure 2 is then

related to the relative equality in information. Figure 3 provides initial evidence that the quality of the weights depends to some degree on the amount and location of test information.

Example 2

For Example 2, the effect of disparate test section information was considered when both sections contain binary items fit with 2PL models. The items in both sections have thresholds centered at $\theta = 0$, but the discrimination parameters for Section 1 are one unit higher for each item (Table 6). Because item information is largely affected by the discrimination parameter, one would expect better score precision when the section containing more information is more heavily weighted.

Table 6. Example 2: IRT parameters for two 2PL test sections

Item	*a*	*b*
Section 1		
MC_1	1.75	-0.50
MC_2	2.00	-0.25
MC_3	2.25	0.25
MC_4	2.50	0.50
Section 2		
MC_1	0.75	-0.50
MC_2	1.00	-0.25
MC_3	1.25	0.25
MC_4	1.50	0.50

The extended recursive algorithm was again iterated over 21 sets of weights and the resulting marginal reliabilities are graphically shown in Figure 4. Unlike Example 1, Figure 4 illustrates that optimal reliability occurs when Section 1 receives nearly twice as much weight as Section 2 (weights of 0.65 and 0.35, for Sections 1 and 2, respectively). While the optimal reliability is 0.74, for a wide range of weights the reliability is greater than 0.70

(including weights from [0.4 and 0.6] to [0.9 and 0.1]). For the majority of weights, the scores are more precise when Section 2 is more heavily weighted.

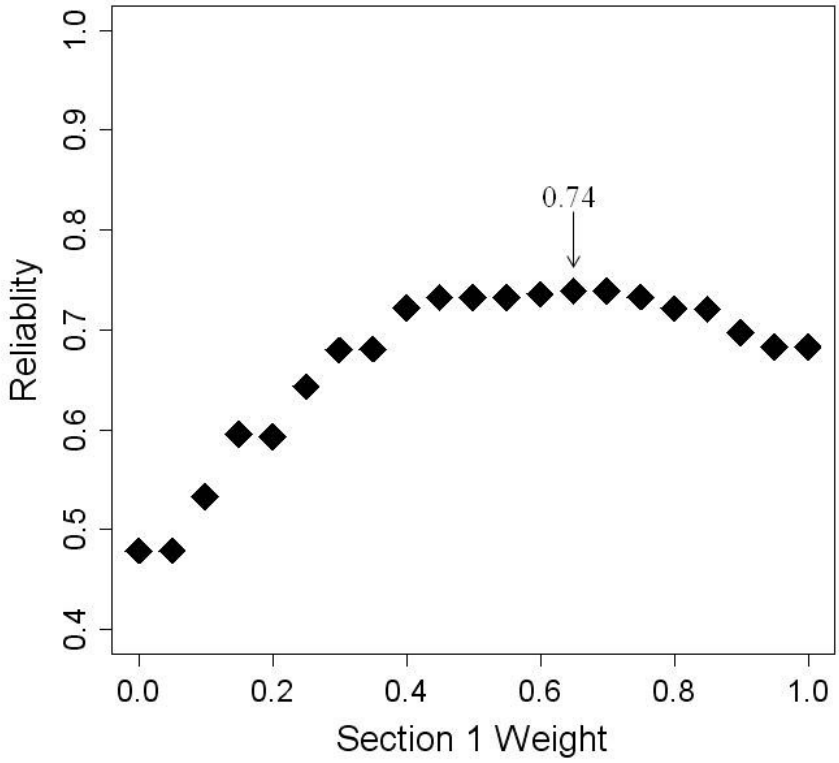

Figure 4. Example 2: Reliabilities for two 2PL test sections

Example 3

Example 3 considers the effects of weighting test sections comprising binary items fit with 2PL and 3PL models. The 3PL model (Birnbaum, 1968) may be considered an extension of the 2PL:

$$T(u_i = 1 \mid \theta) = g_i + \frac{(1 - g_i)}{1 + \exp[-a_i(\theta - b_i)]}. \tag{13}$$

The probability of a correct response 1 to item *i* given the level of the latent variable, θ, is a function of the discrimination power of the item (a_i), the difficulty of the item (b_i), and the degree of guessing (g_i). The lower asymptote, g_i, is the probability of correctly answering an item at an infinitely low level of ability. There are differences in the meanings of the other parameters when $g_i > 0$ that result from accounting for guessing in relating the underlying latent variable to the item responses. For example, b_i in the 2PL is the location on θ halfway between the probabilities of 0.0 and 1.0 for correct response; however, in the 3PL, b_i marks the location on θ halfway between the probabilities of the lower asymptote, g_i, and 1.0. Given these differences, after substituting equation 13 into equation 7, computing EAPs and SDs from weighted summed scores remains unchanged.

For Example 3, Section 1 contains 2PL items while Section 2 contains 3PL items where $g_i = .25$. Discrimination and threshold parameters are equal across test sections (see Table 7 and graphically in Figure 5).

Table 7. Example 3: IRT parameters for test sections with 3PL and 2PL items

Item	*a*	*b*	*g*
Section 1			
MC_1	1.75	-0.50	
MC_2	2.00	-0.25	
MC_3	2.25	0.25	
MC_4	2.50	0.50	
Section 2			
MC_1	1.75	-0.50	0.25
MC_2	2.00	-0.25	0.25
MC_3	2.25	0.25	0.25
MC_4	2.50	0.50	0.25

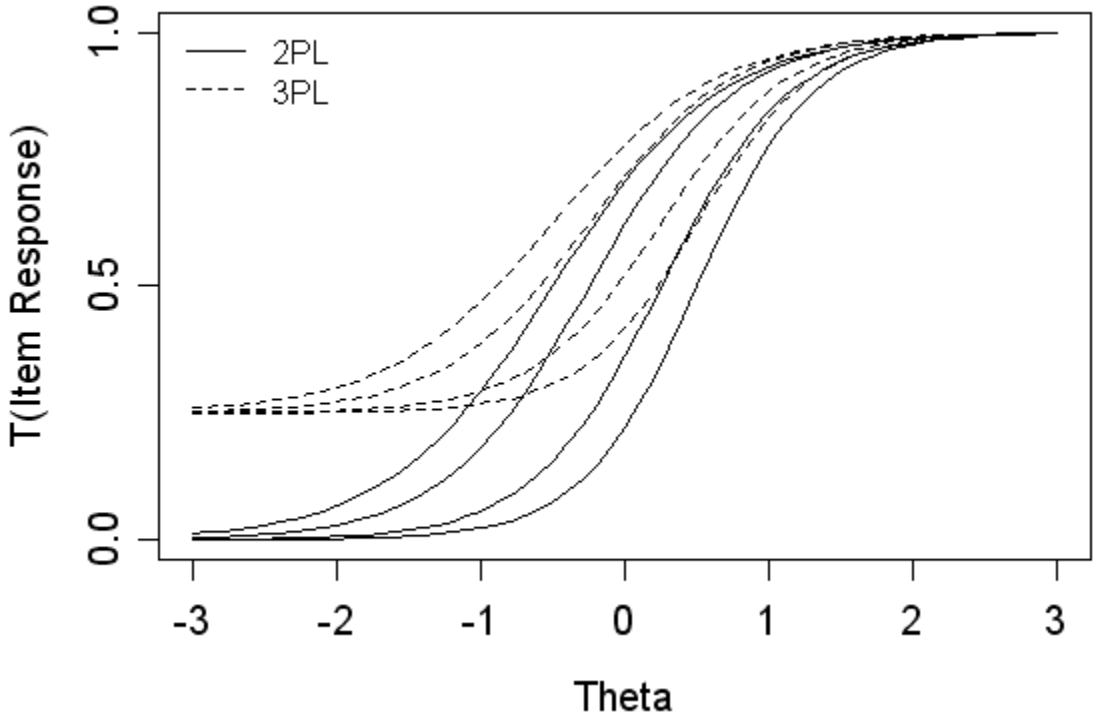

Figure 5. Example 3: Trace lines for 2PL and 3PL items

The extended recursive algorithm was again iterated over 21 sets of weights and the resulting marginal reliabilities are graphically shown in Figure 6. For these models the optimal reliability (0.74) occurs when Section 1 (2PL items) is weighted by 0.6 units and Section 2 (3PL items) by 0.4 units. The 2PL section has an average reliability of 0.68, which is exceeded by all other weight combinations except those which weight the 3PL items three times (or more) greater than the 2PL items. This disparity indicates that after controlling for differences discrimination and threshold parameters, reliability is improved when items that account for guessing receive less weight.

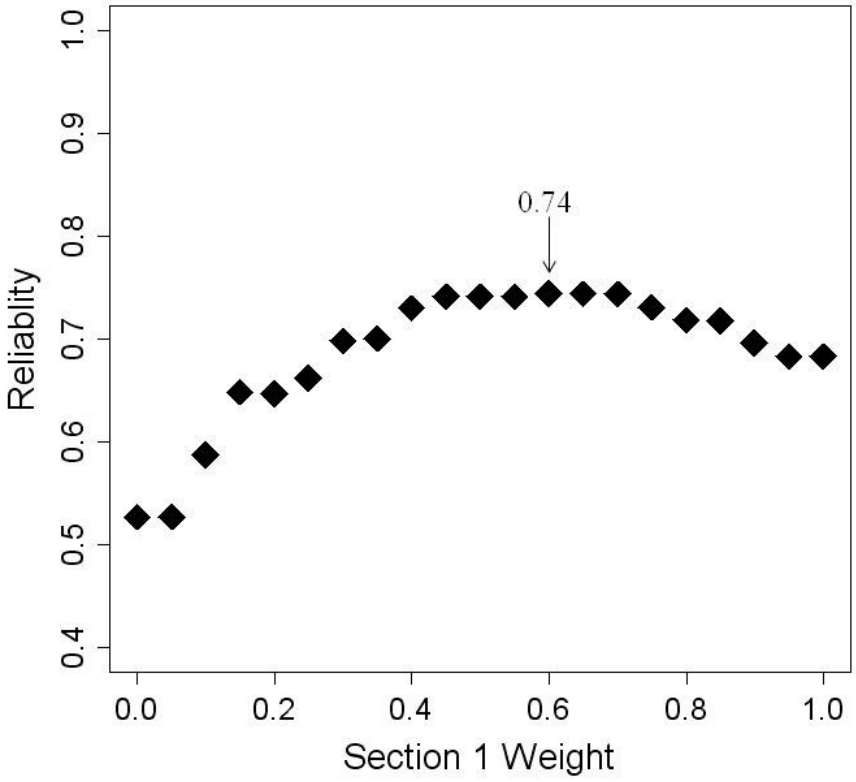

Figure 6. Example 3: Reliabilities for test sections

with 2PL and 3PL items

This finding suggests that items fit with 2PL model have more information than when fit with 3PL model. Item information for the 3PL is:

$$\sum_{1}^{n} a_i^2 T_i(\theta)[(1-T_i)\theta]\left[\frac{T_i^*(\theta)}{T_i(\theta)}\right]^2. \qquad (14)$$

The additional term reduces the information by the square of the ratio of correctly responding to an item under the 2PL model ($T_i^*(\theta)$) to that under the 3PL model ($T_i(\theta)$) (Baker & Kim,

2004, chap. 3). When $g_i > 0$, the reduction in information may be thought of as "penalty" for correctly answering an item by guessing. Birnbaum (1968, pp. 464) demonstrates that under the 3PL model the maximum amount of information will always occur to the right of b_i on the ability scale. Both of these concepts are displayed graphically in Figure 7 which illustrates both the decrease and shift in information for the 3PL items used in Example 3.

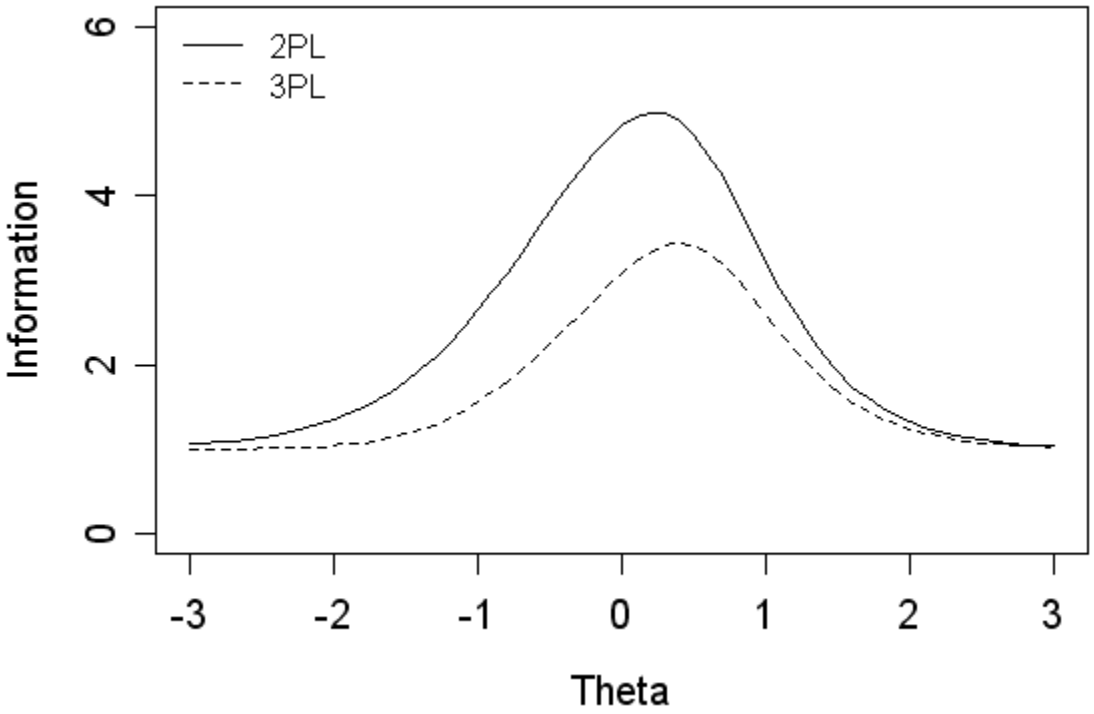

Figure 7. Example 3: Test information for 2PL and 3PL items

Example 4

In the final example two 2PL sections are combined. In Section 1 the thresholds are centered at $\theta = 0$, and for Section 2 the thresholds are uniformly 2 standard deviations higher (see Table 8). For both sections the discrimination parameters are equal, so that the models have equal total test information, but the maximum amount of information for Section 1 is

centered at approximately θ = 0 (not exactly θ = 0 because the discrimination parameters increase as the thresholds increase), and for Section 2 at approximately θ = 2 (see Figure 8).

Table 8. Example 4: IRT parameters for two 2PL test sections with different locations of information

Item	*a*	*b*
Section 1		
MC_1	1.75	-0.50
MC_2	2.00	-0.25
MC_3	2.25	0.25
MC_4	2.50	0.50
Section 2		
MC_1	1.75	1.50
MC_2	2.00	1.75
MC_3	2.25	2.25
MC_4	2.50	2.50

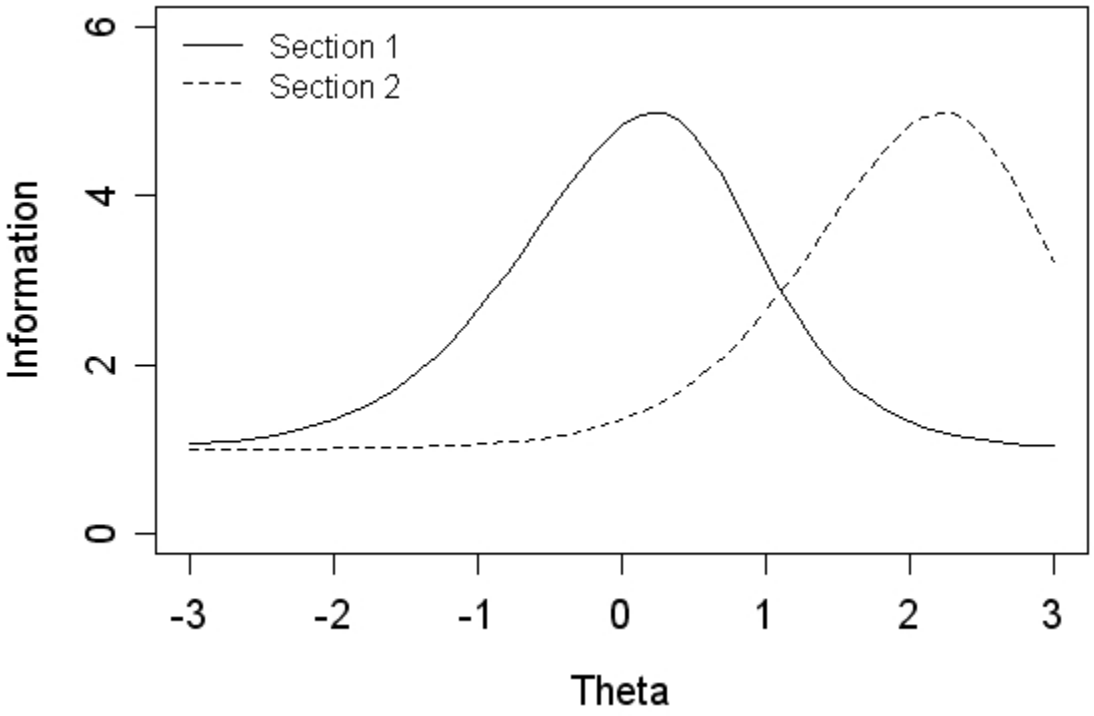

Figure 8. Example 4: Test information for two 2PL sections

Results for Example 4 indicate the location of test information affects the relation between weighting and marginal reliability. As Figure 9 illustrates, the optimal weights occur when Sections 1 and 2 receive approximately equal weight (i.e., weights of [0.45 and 0.55], [0.5 and 0.5], and [0.55 and 0.45]). The degree to which Section 1 received more weight than Section 2 had little impact on the marginal reliability (equally weighted sections were .06 more reliable than the reliability of Section 1 alone). When Section 2 received the majority of weight (i.e., 4 times the weight of Section 1) the reliability of the test is lower than the reliability of Section 1. For the extreme case where Section 2 receives all the weight and Section 1 receives none, the reliability is substantially lower ($\bar{\rho} = 0.35$). This is the result of administering 4 difficult items where the most probable summed score, 0 (79% of examinees), is located where there is little information, or large score variance, and as equation 11 indicates, this greatly increases the average error variance.

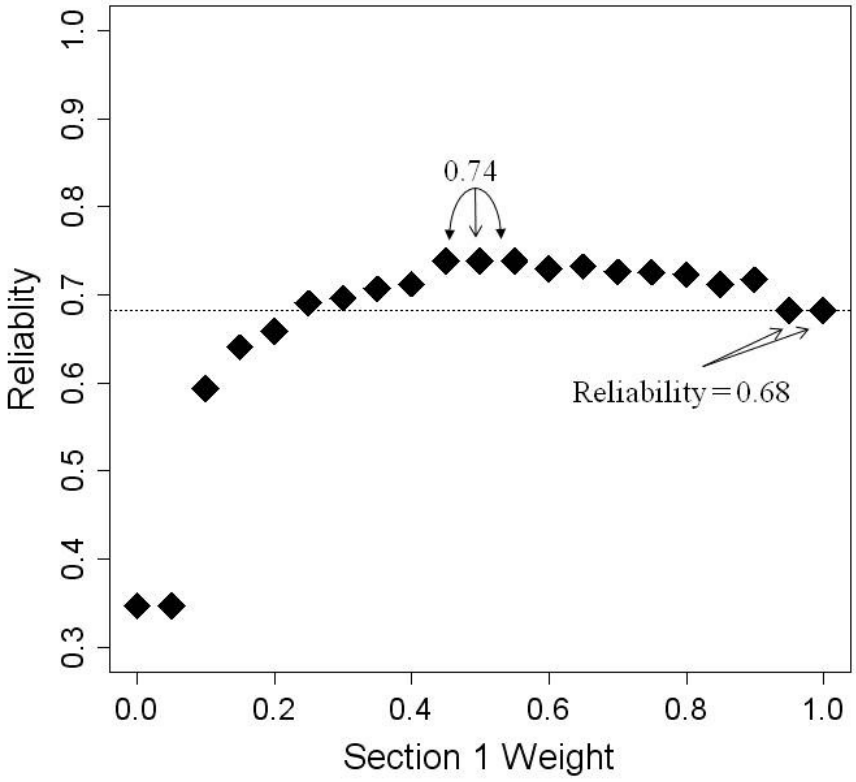

Figure 9. Reliabilities for two 2PL test sections

with different locations of information

An Empirical Example

As an illustration of weighting test sections, the generalized recursive algorithm was used to compute IRT scale scores from weighted summed scores for a large-scale achievement test comprised of 98 binary items and 4 polytomous response items with 11 response options (0 to 10). The 3PL model was fit to the binary items and the GRM to the graded response items (see Appendix I for item parameters and Figure 10 for an illustration of test information). The generalized recursive algorithm was iterated over a range of weights to illustrate the effects of a variety of weight combinations on marginal reliability.

Unlike previous examples, where sections contained the same number of points, here the MC section is explicitly weighted by containing more points than the CR section (98 and 40 points, respectively). It is useful in situations where there is a difference in the number of points for each test section to consider *relative* weights. For this test, relative weights of 0.5 and 0.5 are equivalent to using unit-weights to score each section, and reflect the overall difference in points across sections. If, for example, test officials require that the MC and CR sections be equally weighted, that is, that they generate the same maximum summed score, then relative weights of approximately 0.289 and 0.711 should be used. On a unit-metric, these weights correspond to weighting the MC section by 0.41 units and the CR section by 1.0, this would yield approximately the same maximum summed scores for each section (40 points each).

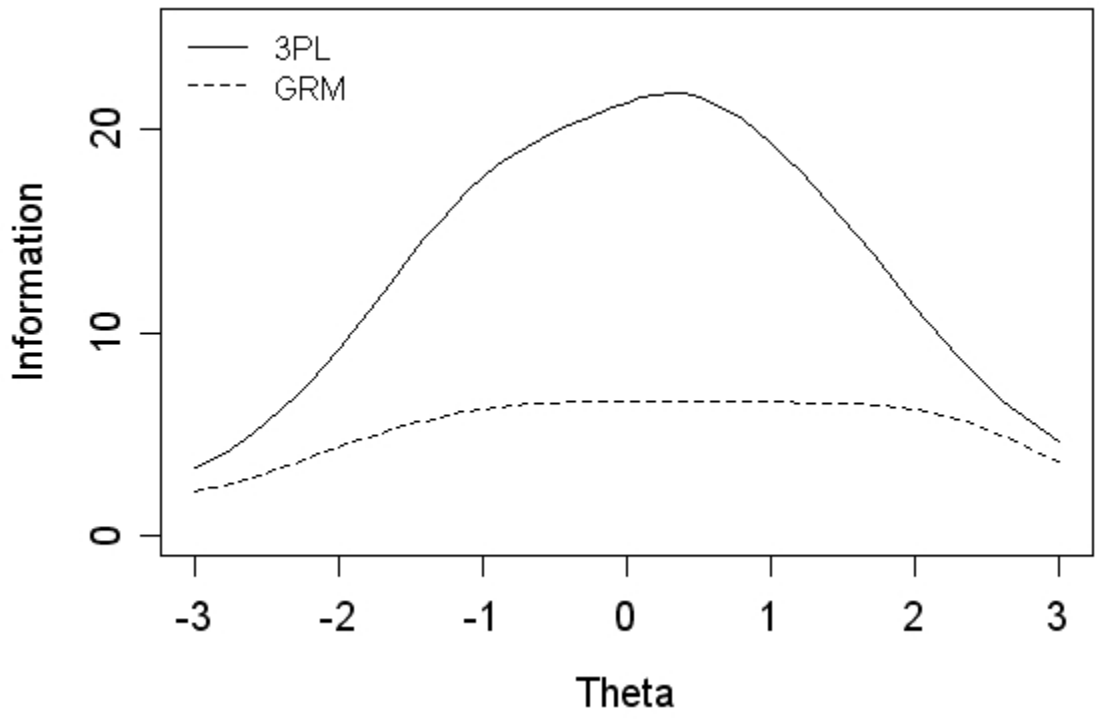

Figure 10. Test information for 98 MC items and 4 CR items

The generalized recursive algorithm was iterated 51 times in increments of 0.02 to compute the marginal reliability for a variety of potential relative weights (see Figure 11). If the test is unit-weighted, that is, when the relative weights are 0.5:0.5, the marginal reliability is 0.9465. When the test is weighted such that each section contributes equally (relative weights of 0.289:0.711), the marginal reliability is 0.9264. When optimal weights are used (relative weights of 0.59:0.41), reliability improves slightly to 0.9483 (see Appendix II for example score reporting tables for the optimal weights).

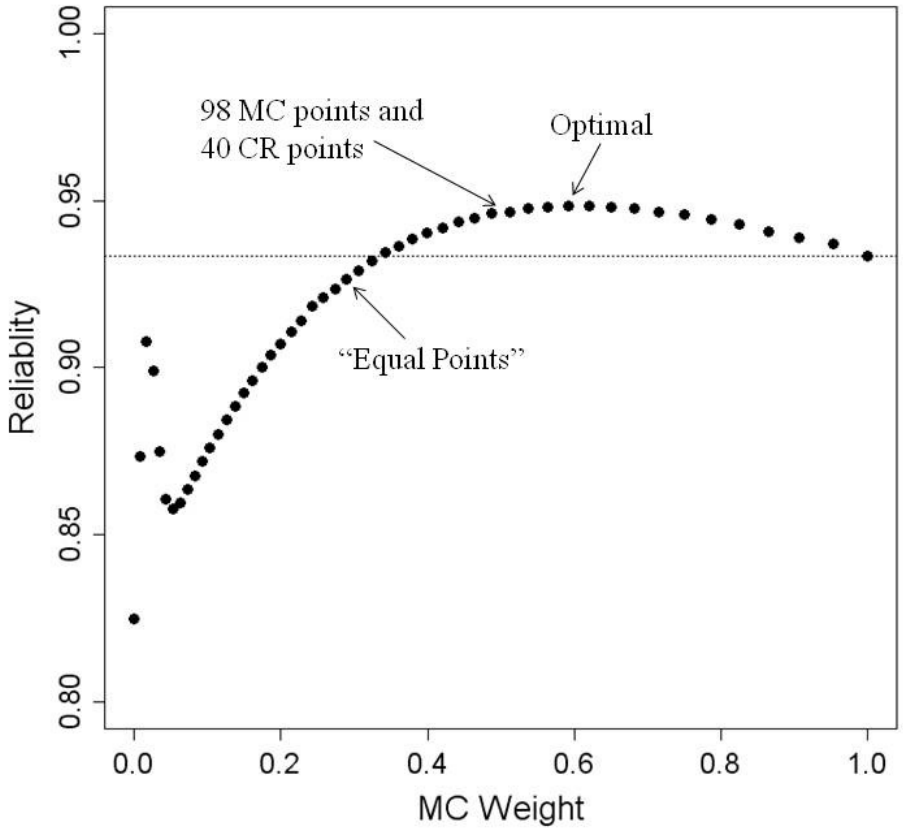

Figure 11. Reliabilities for 51 separate weights for

98 MC items and 4 CR items

For this test there are many sets of weights that produce reliable scores. Using Wainer and Thissen's (2001, chap. 2) concept of weighting to ensure that composite reliability is at least equal to the reliability of most reliable section, a range of relative weights between 1.00:0.00 and 0.34:0.66, provide reliability equal to or greater than the MC section reliability of 0.9334. Within this range it appears to matter very little which weights are chosen. The optimal reliability is only 0.0131 greater than the MC section's reliability, and only 0.0018 greater than the reliability of the unit-weighted composite.

There are also striking anomalies among the values of marginal reliability in Figure 11 when the weight for the 98-point MC section becomes very small, and that of the 40-point CR section becomes large. Under those circumstances, very small changes in the relative weight of the MC section induce much larger changes in the combinations of MC and CR likelihoods that yield scores that are near each integer (recall that weights affect such combinations, as was illustrated using Tables 1 and 2 in Chapter 1). The "smoothing" effect of collapsing the scores to integers that was described in Chapter 2 (p. 16), fails to some degree, because the underlying relation between the EAPs and the weighted summed scores (before collapsing to integers) becomes extremely non-monotonic. This represents more a numerological curiosity than a practical problem, because such extreme weights (the near-equivalent of not-administering a section of the test) would not be used in practice.

CHAPTER 4

THE SELECTION OF OPTIMAL WEIGHTS

The previous section provided a method for comparing the effects of a variety of weighting schemes on marginal reliability. The effectiveness of the weights, however, could not be known prior to iterating the generalized recursive algorithm. Alternatively, rather than relying on comparisons of marginal reliabilities, it would be useful to provide an analytic method which computed optimal weights directly from the item parameters. Such a method would select the optimal weights prior to using the generalized recursive algorithm to compute scale scores from weighted summed scores.

Discimination Parameters as Weights

In IRT models, the item discrimination parameter estimates serves as an implicit weight (Birnbaum, 1968). Because of the association between discrimination parameters and reliability, it may be expected that an optimal set of weights would parallel the item discrimination parameters.

If optimal weights are best represented by differences in discrimination parameters between test sections, then for a test comprised of two sections, it might be expected that the ratio between weights for the sections would be related to the ratio between discrimination parameters. To investigate the potential relationship between weights and discrimination parameters, the generalized recursive algorithm was used to compute optimal reliabilities for a test with two sections containing binary items fit with 1PL models. The models had identically spaced threshold parameters, but the discrimination parameters for the first section were 1.5

times greater than for the second section (see Table 9). Because the ratio of the slopes (3.0 and 2.0) corresponds to the ratio of 0.6 and 0.4, if discrimination parameters best reflect optimal weights, one would expect the best set of weights to be 0.6 and 0.4.

Table 9. IRT parameters for two 1PL test sections
with unequal discrimination parameters

Item	a	b_1
Section A		
MC_1	3.0	-0.50
MC_2	3.0	-0.25
MC_3	3.0	0.25
MC_4	3.0	0.50
Section B		
MC_1	2.0	-0.50
MC_2	2.0	-0.25
MC_3	2.0	0.25
MC_4	2.0	0.50

After computing the marginal reliabilities across a range of weights using the generalized recursive algorithm, the optimal weights for this test were 0.6 and 0.4, which is in agreement with the difference in the magntidue of the slopes and supports the rationale of using differences in slopes to aid in selecting the best weights.

While the previous example suggests weighting test sections based on differences in the magnitude of discrimination parameters, the IRT models used in the example were identical (1PL and 1PL). To consider IRT model-specific effects, in the next example, four binary items were fit with a 1PL model and two polytomous response items (with response categories 0, 1, 2) were fit with a GRM (see Table 10). Notably, the discrimination parameters for these models were identical, and the threshold parameters were selected to provide nearly equal locations of information for both models. Given that 1PL and 2PL models are special cases of the GRM for

binary items (Thissen & Steinberg, 1986), and based on findings from the previous example, it might be expected that that the optimal weights for these models would be equal.

Table 10. IRT parameters for test sections with

1PL and GRM items with equal discrimination parameters

Item	a	b_1	b_2
MC_1	2.0	-0.50	
MC_2	2.0	-0.25	
MC_3	2.0	0.25	
MC_4	2.0	0.50	
CR_1	2.0	-0.50	0.00
CR_2	2.0	0.00	0.50

However, after iterating the extendend recursive algorithm, the optimal weights for this test are 0.65 and 0.35, for the MC and CR sections, respectively. This indicates that the MC section should be weighted nearly twice as much as the CR section. A potential explanation for this result is that, though item discrimination was equal for both sections, the amount of test information was greater for the 1PL items (see Figure 12).

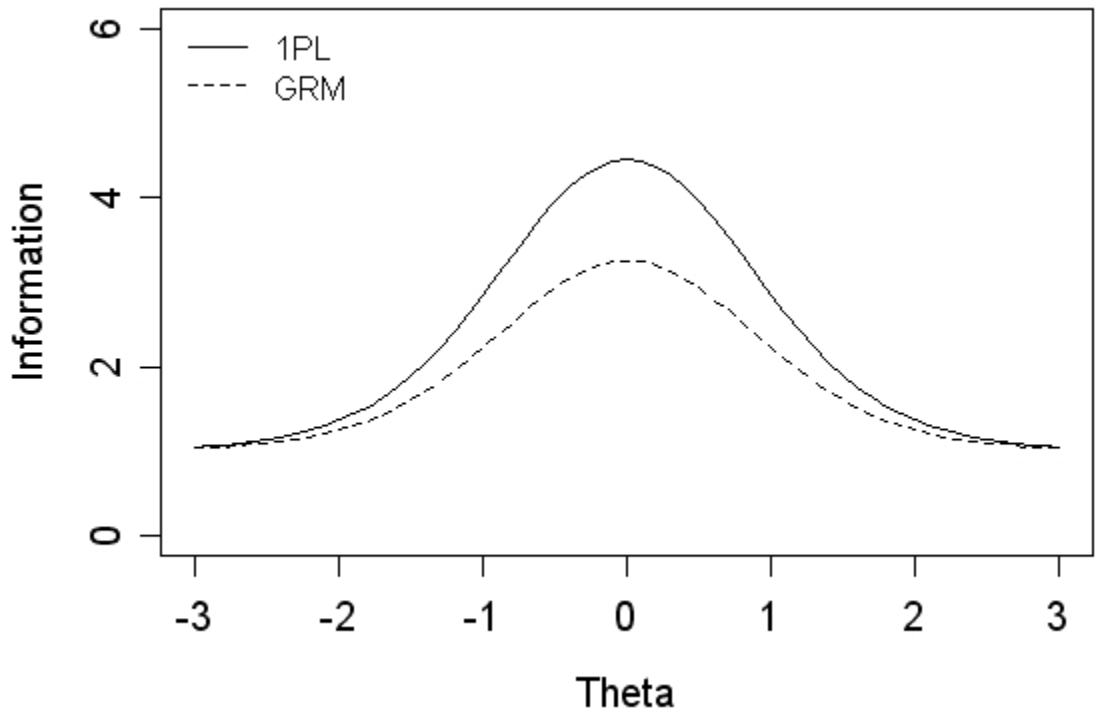

Figure 12. Test information for 1PL and GRM items

Figure 12 illustrates that while the discrimination parameters were equal for these items, test information was greater for the 1PL items, indicating that, in general, optimal weights are not a simple function of discrimination parameters.

Weighting Based on Test Information

To understand the relationships among weights, test information and marginal reliability, selecting weighting schemes based on test information was investigated. Findings from the previous example suggest that marginal reliability may improve when the weights reflect differences in test section information. However, as an added complication, unlike CTT where Cronbach's alpha assumes that all scores have the same standardard error, IRT scale scores vary in precision along θ, and thus no single statistic can exactly reflect the cumulative amount of information from a set of items.

To provide a summary estimate of the overall information for a set of items, average information, weighted by the populatoin distribution, was computed:

$$\bar{I}(\theta) \approx \sum I_i \phi(\theta) d\theta, \qquad 15$$

where average information ($\bar{I}$) is approximately the integral of the product of test information and the population distribution, which is equivalent to multiplying test information by the normal distribution at each point on the θ, and taking the sum. This procedure weights test sections with information located near the center of the distribution and penalizes sections where the information is located near the tails of the distribution. After computing average information for each section, the weights for a test with two sections may be obtained using the following:

$$W_A = \frac{\overline{I_A} \big/ \overline{I_B}}{\left(\overline{I_A} \big/ \overline{I_B}\right) + 1} \text{ and } W_B = 1 - W_A, \qquad 16$$

where W_A and W_B are weights that have been rescaled from the proportion of average information for Sections A and B. Because the average information-based weights are on a metric which sums to 1.0, they are comparable to the values used to weight test sections. If optimal weights can be computed based on the magnitude and location of information, then Eq. 16 should provide a set of weights using average test information which are equal to the optimal weights determined by the extended recursive algorithm.

To consider the efficacy of using average information as a method to analytically obtain optimal weights, in the next example, two sections comprising eight binary items were fit with 1PL models. While the location of information was the same for each section, the 1PL items in

the first section generated more information than the 1PL items in the second section (see Table 11 and graphically in Figure 13).

Table 11. IRT parameters for two test sections with 1 PL items with unequal test information

Item	a	b_1
Section A		
MC_1	3.0	-0.50
MC_2	3.0	-0.25
MC_3	3.0	0.25
MC_4	3.0	0.50
Section B		
MC_1	2.0	-0.50
MC_2	2.0	-0.25
MC_3	2.0	0.25
MC_4	2.0	0.50

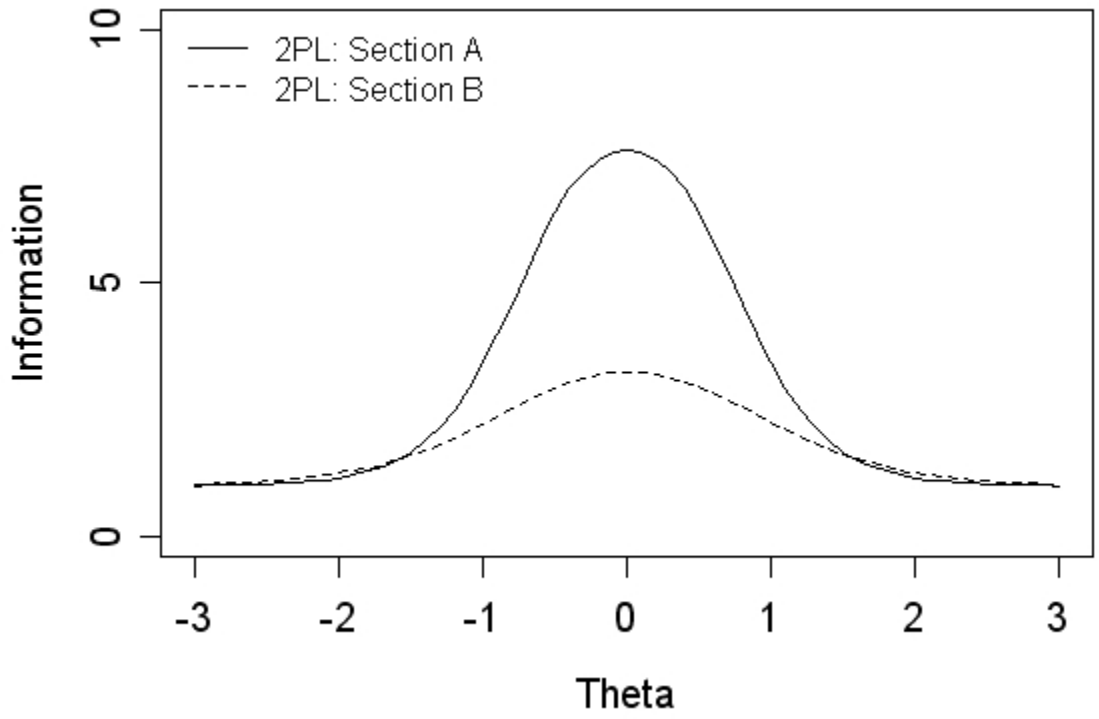

Figure 13. Two test information functions that are similar in location but dissimilar in magnitude.

The average information for the first and second 1PL sections was 3.899 and 2.313, respectively. Using Eqs. 16, the average information for these sections suggests weights of 0.63 and 0.37. After iterating the generalized recursive algorithm, the optimal weights, 0.60 and 0.40, closely reflect the weights obtained using average information.

Given the potential utility of average information in selecting optimal weights, next, rather than changing the magnitude of information, the location of information for the first set of items was shifted 2.5 standard deviations to the right for each 1PL item (see Table 12 and graphically in Figure 14). While the magnitude of information is greater for the first section, the location of information for that set of items is higher on the θ scale.

Table 12. IRT parameters for test sections with 1PL items
with unequal test information and unequal location

Item	a	b_1
Section A		
MC_1	3.0	2.00
MC_2	3.0	2.25
MC_3	3.0	2.75
MC_4	3.0	3.00
Section B		
MC_1	2.0	-0.50
MC_2	2.0	-0.25
MC_3	2.0	0.25
MC_4	2.0	0.50

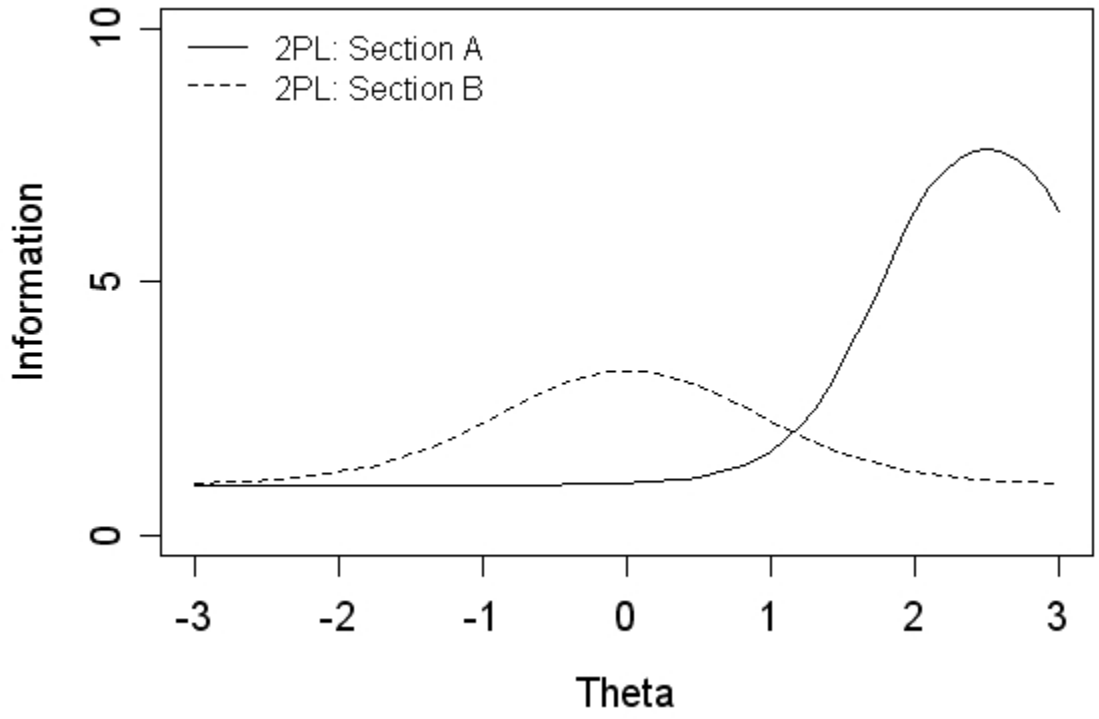

Figure 14. Two test information functions that are dissimilar in location and magnitude.

For these sections, average information suggests weights of 0.18 and 0.82, for the first and second sections, respectively. However, the generalized recursive algorithm indicates that optimal reliability occurs when the section weights are equal (or nearly so) (i.e., 0.45/0.55, 0.50/0.50, and 0.55/0.45). In a sense, the greater magnitude of information in the first section is counter-balanced by its offset location, resulting in optimal weights that are unit-weights. While average information accounts for a shift in the location of test precision, it seems that multiplying by the population distribution may over-penalize models for which the majority of information occurs away from the center of the distribution.

These examples suggest that reliability is improved when weights generally reflect item discrimination or test section information, however, it does not appear that either of these

methods accurately predict optimal weights in any simple way. Specifically, selecting weights based on item discrimination appears to succeed when comparing IRT models for the same number of response categories, but fails when comparing tests comprised of mixed item types. Selecting weights based on average test information is useful when the information occurs near the center of the distribution, however, the computation of average information devalues sections where the information is located in either tail of the distribution. It may be that these procedures are over simplifications of known complexities regarding IRT score combinations.

The Relation between Section and IRT-Optimal Weights

Thus far we have only considered weights applied to all scores on each test section. As an alternative to these methods, Thissen, Nelson, and Swygert (2001, chap. 8) provide a technique for computing approximations of scale scores for linear combinations of component scores. The method uses a somewhat different type of weight to compute approximated scale scores. After computing the EAPs for the sum scores in each component, the component EAPs are combined by weighting each component score by that particular score's precision. The weight used is the inverse of the variance associated with each scale score computed from summed score:

$$w_x = 1/\sigma^2(\theta \mid x) \qquad 17$$

The inverse of the score variance serves to weight each score combination. The intended use of weights in this sense is only to provide a linear combination of scale scores from separate components which approximate the scale scores computed from patterns of summed scores. However, as a byproduct of this method, the weights provide additional information regarding the precision of component score combinations. The ratio between the weights for each component score indicates, over a range of scores, where a particular component provides the

most precise score estimates. It is important to note that their weights vary across scores within sections.

Optimal weights, as previosly described in chapter 3, are simplifications and approximations to the IRT score-specific weights used in chapter 8 of *Test Scoring*. Thissen et al. (2001, chap. 8) note that no single combination of weights (i.e., constant weights) will be best because score-specific weights may vary greatly across scores. However, the set (or sets) of weights that produce optimal reliability are likely the best constant-weight approximations to the score-specific weights. Score-specific weights provide information about the differences in score precision for each possible component score combination, while optimal weights provide a single summary value for all possible score combinations.

What remains are two general approaches for weighting in IRT. The procedures described here provide a technique for the analysis of a variety of test section weights from an IRT perspective, while Chapter 8 of *Test Scoring* descibes score-specific weights which are useful in understanding differences in the precision of combinations of scores across test sections. Both techniques should generally agree about how test sections should best be weighted, though the generality of the method introduced here allows for the computation of IRT scale scores for tests comprising multiple test sections with any set of user-defined weights.

Appendix I:

Item parameters for 98 multiple choice items

and 4 graded response items

CR Item	a	b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8	b_9	b_{10}
1	2.27	-1.87	-1.37	-0.90	-0.47	-0.04	0.36	0.76	1.17	1.64	2.28
2	1.39	-3.78	-2.77	-1.96	-1.30	-0.64	0.15	0.87	1.61	2.57	3.75
3	2.14	-1.63	-1.13	-0.75	-0.40	-0.03	0.32	0.63	1.01	1.42	1.92
4	2.36	-0.89	-0.44	-0.09	0.21	0.51	0.82	1.16	1.58	1.97	2.45

MC Item	a	b	g
1	1.14	-2.63	0.21
2	0.62	-1.35	0.21
3	1.44	-1.72	0.27
4	1.41	-0.30	0.11
5	1.18	-1.71	0.18
6	0.79	-0.80	0.19
7	0.65	-1.66	0.17
8	1.42	-0.59	0.13
9	0.80	-0.96	0.13
10	1.02	-0.64	0.10
11	0.62	0.37	0.28
12	0.58	-0.55	0.12
13	1.54	-0.58	0.19
14	0.99	0.29	0.10
15	2.10	-0.80	0.36
16	1.99	-0.91	0.11
17	1.27	-0.11	0.15
18	1.16	-0.91	0.40
19	0.34	-0.89	0.20
20	1.87	-0.97	0.21
21	1.07	-0.31	0.20
22	0.93	-0.30	0.21
23	1.43	0.20	0.17
24	1.79	-0.08	0.31
25	1.91	1.20	0.30
26	0.97	1.25	0.17
27	1.23	1.75	0.15
28	1.81	0.57	0.18
29	1.75	-0.24	0.16
30	1.62	1.12	0.13
31	1.24	-0.68	0.12
32	1.37	0.23	0.19
33	0.86	1.04	0.13

(Continued)

MC Item	*a*	*b*	*g*
34	0.69	-1.34	0.17
35	1.81	1.45	0.16
36	1.72	1.54	0.22
37	1.66	0.97	0.25
38	1.05	0.04	0.13
39	1.52	1.68	0.07
40	1.03	1.40	0.15
41	1.46	0.91	0.11
42	0.85	0.11	0.13
43	1.23	0.37	0.21
44	1.33	1.20	0.06
45	1.72	1.34	0.12
46	1.62	1.35	0.11
47	1.77	0.79	0.10
48	0.95	0.70	0.12
49	1.38	0.10	0.18
50	1.35	0.13	0.13
51	1.44	-1.28	0.16
52	2.07	0.27	0.28
53	1.16	0.37	0.22
54	1.87	0.42	0.15
55	1.54	0.18	0.16
56	1.02	-1.78	0.17
57	1.30	-1.24	0.16
58	1.83	-1.37	0.14
59	0.84	0.37	0.23
60	1.45	-1.24	0.31
61	1.40	-0.03	0.16
62	1.03	-0.26	0.15
63	1.23	0.89	0.26
64	0.61	-2.97	0.14
65	1.51	-0.35	0.18
66	1.12	-0.29	0.22
67	1.29	0.08	0.16
68	1.52	-0.24	0.12
69	2.31	0.34	0.18
70	2.59	0.40	0.18
71	2.02	1.58	0.32
72	1.51	-1.72	0.15
73	1.39	-1.69	0.16
74	1.27	-1.04	0.16
75	1.19	-1.17	0.16
76	1.47	-0.85	0.19

(Continued)

MC Item	*a*	*b*	*g*
77	1.39	-0.73	0.10
78	1.42	-0.05	0.08
79	1.13	-1.53	0.17
80	1.13	-1.25	0.15
81	1.72	-1.85	0.11
82	0.93	-2.27	0.15
83	0.65	0.42	0.15
84	0.92	-1.30	0.16
85	1.03	-1.50	0.12
86	1.17	-1.17	0.20
87	1.09	0.43	0.14
88	0.95	-1.14	0.18
89	0.40	-0.59	0.14
90	1.59	-1.24	0.14
91	1.40	1.33	0.39
92	2.09	-1.35	0.19
93	1.22	-0.63	0.18
94	1.29	-1.50	0.11
95	1.18	1.51	0.07
96	1.74	0.76	0.21
97	1.55	-0.92	0.10
98	1.56	-1.00	0.11

Appendix II:

IRT scale scores for integer scores computed from the optimal weights

of 98 multiple choice items and 4 graded response items

Integer Score	EAP	SD	Integer Score	EAP	SD	Integer Score	EAP	SD
0	-3.94	0.48	37	-1.84	0.31	74	-0.26	0.21
1	-3.89	0.48	38	-1.77	0.30	75	-0.22	0.21
2	-3.85	0.48	39	-1.72	0.29	76	-0.19	0.21
3	-3.81	0.48	40	-1.67	0.29	77	-0.15	0.21
4	-3.76	0.48	41	-1.62	0.28	78	-0.11	0.21
5	-3.71	0.48	42	-1.56	0.28	79	-0.08	0.21
6	-3.50	0.44	43	-1.52	0.27	80	-0.05	0.21
7	-3.59	0.48	44	-1.48	0.27	81	-0.01	0.21
8	-3.54	0.48	45	-1.42	0.26	82	0.03	0.21
9	-3.56	0.49	46	-1.38	0.26	83	0.06	0.21
10	-3.48	0.48	47	-1.33	0.26	84	0.10	0.21
11	-3.42	0.48	48	-1.29	0.25	85	0.14	0.21
12	-3.40	0.48	49	-1.24	0.25	86	0.18	0.21
13	-3.33	0.48	50	-1.20	0.25	87	0.21	0.20
14	-3.26	0.48	51	-1.16	0.24	88	0.25	0.20
15	-3.20	0.47	52	-1.11	0.24	89	0.29	0.21
16	-3.15	0.47	53	-1.07	0.24	90	0.32	0.20
17	-3.00	0.43	54	-1.03	0.23	91	0.36	0.20
18	-3.01	0.46	55	-0.99	0.23	92	0.40	0.21
19	-2.96	0.45	56	-0.94	0.23	93	0.43	0.21
20	-2.89	0.46	57	-0.91	0.23	94	0.47	0.20
21	-2.83	0.44	58	-0.87	0.23	95	0.51	0.21
22	-2.76	0.44	59	-0.83	0.23	96	0.55	0.21
23	-2.70	0.43	60	-0.79	0.22	97	0.58	0.21
24	-2.65	0.43	61	-0.75	0.22	98	0.63	0.21
25	-2.57	0.41	62	-0.71	0.22	99	0.67	0.21
26	-2.50	0.40	63	-0.67	0.22	100	0.70	0.21
27	-2.45	0.40	64	-0.63	0.22	101	0.74	0.21
28	-2.35	0.37	65	-0.60	0.22	102	0.79	0.21
29	-2.31	0.38	66	-0.56	0.22	103	0.82	0.21
30	-2.24	0.37	67	-0.52	0.21	104	0.86	0.21
31	-2.17	0.36	68	-0.48	0.21	105	0.91	0.21
32	-2.12	0.35	69	-0.45	0.21	106	0.95	0.21
33	-2.06	0.34	70	-0.41	0.21	107	0.99	0.21
34	-2.01	0.33	71	-0.37	0.21	108	1.04	0.21
35	-1.93	0.32	72	-0.33	0.21	109	1.08	0.22
36	-1.89	0.32	73	-0.30	0.21	110	1.12	0.22

(Continued)

Integer Score	EAP	SD
111	1.17	0.22
112	1.22	0.22
113	1.27	0.22
114	1.31	0.22
115	1.36	0.23
116	1.41	0.23
117	1.46	0.23
118	1.51	0.23
119	1.57	0.24
120	1.62	0.24
121	1.68	0.24
122	1.75	0.25
123	1.81	0.25
124	1.87	0.25
125	1.94	0.26
126	2.01	0.27
127	2.08	0.27
128	2.17	0.28
129	2.25	0.29
130	2.35	0.30
131	2.44	0.31
132	2.54	0.32
133	2.67	0.34
134	2.78	0.36
135	2.93	0.38
136	3.10	0.41
137	3.30	0.44
138	3.51	0.47

REFERENCES

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord & M. R. Novick, *Statistical theories of mental test scores*. Reading, Mass.: Addison-Wesley, 1968.

Baker, F. B. & Kim, S. H. (2004). *Item Response Theory*: *Parameter Estimation Techniques*. New York, NY: Marcel Dekker, Inc.

Cohen, P., Cohen, J., West, S. G., & Aiken, L. S. (2002). *Applied Multiple Regression/Correlation analysis for the Behavioral Sciences* (3rd Ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement*, *21*, 347-360.

Kane, M. & Case, S. M. (2004). The reliability and validity of weighted composite scores. *Applied Measurement in Education*, *17*, 221-240.

Lord, F. M. & Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Welsley Publishing Company.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings." *Applied Psychology Measurement*, *8*, 453-461.

Lukhele, R., & Sireci, G. (1995). "Using IRT to combine multiple-choice and free-response sections of a test onto a common scale using a priori weights. Annual Meeting of the National Council on Measurement in Education, San Francisco, CA, April 19-21, 1995.

Peel, E. A. (1948). Prediction of a complex criterion and battery reliability. *British Journal of Psychology, Statistical Section*, *1*, 84-94.

Rudner, L. M. (2001). Informed test component weighting. *Educational Measurement*, Spring, 16 – 19.

Samejima, F. (1974). Normal ogive model on the continuous response level while in the multidimensional latent space. *Psychometrika, 39*, 111-121

Sykes, R. C. & Hou L. (2003). Weighting constructed-response items in IRT-based exams. *Applied Measurement in Education*, *16*, 257 – 275.

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). Hillsdale, NJ: Lawrence Erlbaum Associates.

Thissen, D., Nelson, L., & Swygert, K.A. (2001). Item response theory applied to combinations of multiple-choice and constructed-response items and approximation methods for scale scores. In D. Thissen & H. Wainer (Eds), *Test Scoring* (pp. 293-341). Hillsdale, NJ: Lawrence Erlbaum Associates.

Thissen, D., & Orlando, M. (2001). Item response theory for items scored in two categories. In D. Thissen & H. Wainer (Eds), *Test Scoring* (pp. 73-140). Hillsdale, NJ: Lawrence Erlbaum Associates.

Thissen, D., Pommerich, M., Billeaud, K., & Williams, V. S. L. (1995). Item response theory for scores on tests including polytomous items with ordered responses. *Applied Psychological Measurement*, *19*, 39 – 49.

Thissen, D. & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, *51*, 567-577.

Thissen, D., Wainer, H., & Wang, X.B. (1994). Are tests comprising both multiple-choice and free-response items necessarily less unidimensional than multiple-choice tests? An analysis of two tests. *Journal of Educational Measurement, 31*, 113-123.

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* (2), 103 – 118.

Wainer, H. & Thissen, D. (2001). True score theory: The traditional method. In D. Thissen & H. Wainer (Eds), *Test Scoring* (pp. 23-72). Hillsdale, NJ: Lawrence Erlbaum Associates.