Item Response Theory for Scores on Tests Including Polytomous Items with Ordered Responses

David Thissen, University of North Carolina at Chapel Hill
Mary Pommerich, American College Testing
Kathleen Billeaud, University of North Carolina at Chapel Hill
Valerie S. L. Williams, National Institute of Statistical Sciences

Item response theory (IRT) provides procedures for scoring tests including any combination of rated constructed-response and keyed multiple-choice items, in that each response pattern is associated with some modal or expected a posteriori estimate of trait level. However, various considerations that frequently arise in large-scale testing make response-pattern scoring an undesirable solution. Methods are described based on IRT that provide scaled scores, or estimates of trait level, for each summed score for rated responses, or for combinations of rated responses and multiple-choice items. These methods may be used to combine the useful scale properties of IRT-based scores with the practical virtues of a scale based on a summed score for each examinee. *Index terms: graded response model, item response theory, ordered responses, polytomous models, scaled scores.*

Item response theory (IRT) provides a score scale that is more useful for many purposes (e.g., for the construction of developmental scales or for the calibration of tests comprising different types of items or exercises) than the summed score, percentage correct, or percentile scales. With the exception of the Rasch family of models for which the summed score is a sufficient statistic for the characterization of the latent variable (θ) (Masters & Wright, 1984; Rasch, 1960), under IRT models each response pattern is usually associated with a unique estimate of θ. These estimates of θ may be used as scaled response pattern scores; they have the advantage that they extract all information available in the item responses, if the model is appropriate for the data. In addition, the IRT model produces estimates of the probability that each response pattern will be observed in a sample from a specified population.

In applied measurement contexts, however, it is often desirable to consider the implications of IRT analysis for summed scores, rather than response patterns, even if the IRT model used is not part of the Rasch family. For example, in a large-scale testing program it may be desirable to tabulate the IRT scaled scores associated with each summed score on operational forms, using item parameter estimates obtained from item tryout data, before the operational forms are administered. In addition, it may be useful to compute model-based estimates of the summed score distribution (e.g., to create percentile tables for use as an interpretive aid for score reporting). Model-based estimates of the summed score distribution also may have value as a statistical diagnostic of the goodness of fit of the IRT model, including the validity of the assumed underlying population distribution.

Many contemporary tests include extended constructed-response items, for which the item scores are ordered categorical ratings provided by raters. In some cases, the constructed-response items comprise the entire test; in other cases, there are multiple-choice items as well. In either case, some total score is often required, combining the ratings of the constructed-response items (and the item scores on the multiple-

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 19, No. 1, March 1995, pp. 39–49

0146-6216/95/010039-11$1.80

choice items, if any are present). Simple summed scores may not be useful in this context, because of the problems associated with the selection of relative weights for the different items and item types, and because the constructed-response items are often on forms of widely varying difficulty. If the collection of items is sufficiently well represented by a unidimensional IRT model, scaled summed scores may be a viable scoring alternative.

IRT for Summed Scores

For any IRT model for items indexed by i with ordered item scores $k = 0, \ldots, K_i$, collected in the vector $\mathbf{k}$, the likelihood for any summed score $j = \Sigma k_i$ is

$$L_j(\theta) = \sum_{\ni j = \Sigma k_i}^{patterns} L(\mathbf{k}|\theta), \tag{1}$$

where the summation is over the response patterns with total score j. The likelihood for each response pattern is

$$L(\mathbf{k}|\theta) = \prod_i T_{k_i}(\theta)\phi(\theta), \tag{2}$$

where $T_{k_i}(\theta)$ is the category response function (CRF) for category k of item i (i.e., the conditional probability of response k to item i given θ) and $\phi(\theta)$ is the population density. Thus, the likelihood for each score is

$$L_j(\theta) = \sum_{\ni j = \Sigma k_i}^{patterns} \prod_i T_{k_i}(\theta)\phi(\theta). \tag{3}$$

Therefore, the probability of each score j is

$$P_j = \int L_j(\theta)\, d\theta, \tag{4}$$

or

$$P_j = \int \sum_{\ni j = \Sigma k_i}^{patterns} L(\mathbf{k}|\theta)\, d\theta, \tag{5}$$

or

$$P_j = \int \sum_{\ni j = \Sigma k_i}^{patterns} \prod_i T_{k_i}(\theta)\phi(\theta)\, d\theta. \tag{6}$$

Given an algorithm to compute the integrand in Equation 1, it is straightforward to compute the average [or expected a posteriori (EAP)] scaled score (Bock & Mislevy, 1982) associated with each score,

$$\mathrm{EAP}(\theta|j = \Sigma k_i) = \frac{\int \theta L_j(\theta) d\theta}{P_j}, \tag{7}$$

and the corresponding standard deviation (SD),

$$\mathrm{SD}(\theta|j = \Sigma k_i) = \left\{ \frac{\int \left[\theta - \mathrm{EAP}(\theta|j = \Sigma k_i)\right]^2 L_j(\theta) d\theta}{P_j} \right\}^{1/2}. \tag{8}$$

The values computed using Equation 7 may be tabulated and used as the IRT scaled-score transformation of the summed scores, and the values of Equation 8 may be used as a standard description of the uncertainty associated with those scaled scores.

The score histogram created using the values of Equation 6 may be used to construct summed-score

percentile tables; if the IRT model fits the data, this can be done accurately using only the item parameters for any group with a known population density. Thus, percentile tables for summed scores can be constructed using item tryout data, before the operational test is administered. This same histogram may also prove useful as a diagnostic statistic for the goodness of fit of the model by comparing the modeled representation of the score distribution to the observed data.

Algorithms for Computing $L_j(\theta)$

Lord (1953) used heuristic procedures to describe the difference between the distribution of summed scores, $L_j(\theta)$, and the underlying distribution of θ, $\phi(\theta)$ (see also Lord & Novick, 1968, pp. 387–392). However, practical calculation of the summed score distribution implied by an IRT model has awaited both contemporary computational power and solutions to the apparently intractable computational problem.

The Brute-Force Method

An exact numerical brute-force evaluation of Equation 6, requiring the computation of $\prod(K_i + 1)$ likelihoods, is possible for a few items; but it is inconceivable for many items. Brute-force may be extended to approximately 20 items by using an algorithm involving the computation of each pattern likelihood from some other previously computed pattern likelihood by a single (list) multiplication; this approach is used in the computer program TESTFACT (Wilson, Wood, & Gibbons, 1991). For binary items, by carefully ordering the computation of the likelihoods for the 2^n patterns (where n is the number of items), such an algorithm can compute all 2^n likelihoods at a computational cost of only a single (list) multiplication for each (Thissen, Pommerich, & Williams, 1993). Nevertheless, due to the exponential computational complexity of this approach, this algorithm cannot be extended to more items regardless of improvements in computational speed.

An Approximation Method

Lord & Novick (1968) stated that "...approximations appear inevitable..." (p. 525), and suggested the use of an approximation to the compound binomial, attributed to Walsh (1963), to compute the likelihood of a summed score for binary items as a function of θ. For n items, this Taylor-series expansion has n terms; however, in practice the first two terms suffice for acceptable accuracy. The two-term version of the approximation is:

$$\sum_{\ni j=\Sigma k_i}^{patterns} \prod_i T_{k_i}(\theta) \cong p_n(j) + \frac{n}{2} VC(j), \tag{9}$$

where

$$p_n(j) = \begin{cases} \binom{n}{j} M^j \left(1 - M^{n-j}\right) & \text{for } j = 0, 1, \ldots, n; \\ 0 & \text{otherwise} \end{cases}, \tag{10}$$

$$C(j) = \sum_{v=0}^{2} (-1)^{v+1} \binom{2}{v} p_{n-2}(j - v), \tag{11}$$

$$V = \frac{1}{n} \sum_i \left[T_{1_i}(\theta) - M\right]^2, \tag{12}$$

and

$$M = \frac{1}{n}\sum_{i} T_{i}(\theta). \tag{13}$$

Yen (1984) used this approximation to develop an algorithm to compute the mode of

$$\sum_{\ni j=\Sigma k_i}^{patterns} \prod_{i} T_{k_i}(\theta) \tag{14}$$

for use as a scaled score for examinees with summed score j on a binary test using the three-parameter logistic model. She reported that the two-term Taylor expansion produced noticeably better results than the one-term solution, which is simply an inverse transformation of the test response function; but the three- and four-term solutions appeared not to add useful precision.

The approximation in Equation 9 also may be substituted for the sum of products in Equations 6, 7, and 8 to compute P_j, $\mathrm{EAP}(\theta|j = \Sigma k_i)$, and $\mathrm{SD}(\theta|j = \Sigma k_i)$. When the results for the two-term approximation were compared to the exact results for one of the 20-item examples used by Yen (1984), the error of approximation was usually less than .001 for $\mathrm{EAP}(\theta|j = \Sigma k_i)$, and $\mathrm{SD}(\theta|j = \Sigma k_i)$ (Thissen et al., 1993). Perfect scores were the exception because the second term of the two-term approximation was 0; for those scores, the approximation was inexact by as much as .05. The error of approximation for P_j was approximately .0001. For practical use in constructing score reporting tables, which usually use no greater precision than tenths of a SD for the scores and their standard errors, and integral values for percentile tables, this degree of precision appears to be sufficient. However, the approximation in Equation 9 is still somewhat computationally burdensome, and no generalization has been offered for items with more than two response categories.

A Recursive Algorithm

The problem of the computational burden is solved by an alternative procedure briefly described by Lord & Wingersky (1984). Abandoning the contention of Lord & Novick (1968) that "...approximation is inevitable..." (p. 525), Lord and Wingersky described a simple recursive algorithm for the computation of

$$L_j^n(\theta) = \sum_{\ni j=\Sigma k_i}^{patterns} \prod_{i} T_{k_i}(\theta) \tag{15}$$

for binary items. The algorithm is based on the distributive law, and generalizes readily to items with any number of response categories.

The generalization follows: Let $i = 0, 1, \ldots, n$ for the items (it is somewhat unusual to index the items from 0 to n for $n + 1$ items; however, in this case the correspondence of that system with the usual practice of indexing the scores from 0, and the common practice of indexing the item response categories from 0, simplifies both the notation and the software implementation); $k = 0, 1, \ldots, K_i$ for the response categories for item i; and $T_{k_i}(\theta)$ be the CRF for category k of item i. In addition, the summed scores for a set of items $[0 \ldots n^*]$ are $j = 0, 1, \ldots, \Sigma_{n^*}(K_i)$ and the likelihood for summed score j for a set of items $[0 \ldots n^*]$ is $L_j^{n^*}(\theta)$; the population distribution is $\phi(\theta)$.

The generalized recursive algorithm is:

Set $n^* = 0$

$L_j^{n^*}(\theta) = T_{jn^*}(\theta)$, for $j = 0, 1, \ldots, K_{n^*}$

Repeat:

For item $n^* + 1$ and scores $j = 0, 1, \ldots, \Sigma_{n^*}(K_i)$,

$$L_{j+k}^{n^*+1}(\theta) = \sum_{k_{n^*+1}} L_j^{n^*}(\theta) T_{k_{n^*+1}}(\theta). \tag{16}$$

Set $n^* = n^* + 1$

Until $n^* = n$.

For a sample from a population with distribution $\phi(\theta)$, the likelihood for score j is

$$L_j(\theta) = L_j^n(\theta)\phi(\theta), \tag{17}$$

and $\mathrm{EAP}(\theta|j = \Sigma k_i)$, $\mathrm{SD}(\theta|j = \Sigma k_i)$, and $P_j(\theta)$ can be computed by integrating $L_j(\theta)$.

No particular parametric form for the CRFs is assumed in the formulation of the recursive algorithm. In the North Carolina testing program, for which some of this system was developed, the three-parameter logistic model is used with binary-scored multiple-choice items and Samejima's (1969) graded model is used for multiple-category rated items. However, in principle, any CRFs could be used, such as the nonparametric kernel smooths described by Ramsay (1991). The algorithm would produce accurate, but meaningless, results if it were used with items for which the responses are not ordered. The results would be meaningless because the response patterns included in any particular summed score would not have likelihoods concentrated near the same values of θ; therefore, such summed-score likelihoods would tend to be very flat with very large SDs.

Nevertheless, the algorithm is completely general. [An implementation for the LISP-STAT computing environment (Tierney, 1990) is available from the author.] For simplicity of programming, it uses rectangular quadrature, or the "repeated midpoint formula" (Stroud, 1974, p. 120), to compute the values of the integrals. Stroud described a number of alternative methods for numerical evaluation of such integrals. Some of the more complex methods, such as Gauss-Hermite quadrature, have often been used in IRT. Stroud (1974) noted that although "often a Gauss formula will be much superior to any other formula with the same number of points... It is not true, however, that a Gauss formula is always the best" (p. 187). For the integration of functions that depend on a large number of unknown parameters, such as those considered here, Stroud recommended that various quadrature methods be compared over a wide variety of possible values of the parameter set to determine the best method. If such a comparison were to be done, it would be very useful for many other applications of IRT, as well as that discussed here.

A Numerical Example

This example is based on three binary items with $a_0 = .5$, $b_0 = -1.0$, $a_1 = 1.0$, $b_1 = 0.0$, $a_2 = 1.5$, $b_2 = 1.0$, and seven quadrature points at $\theta = -3, -2, -1, 0, 1, 2$, and 3. Numerical representations of the item response functions (IRFs) and a number of intermediate results are shown in Table 1. The uppermost section shows the values of the IRFs at the seven values of θ. For $n^* = 0$, there are only two possible scores, 0 and 1, and $L_j^0(\theta)$ is equal to $T_{j0}(\theta)$. Then, as n^* increases and each successive item is used, the likelihood for a score is the sum of the two terms: the product of the likelihood for that score on the preceding items and $T_{0n^*}(\theta)$, and the product of the likelihood for that score minus 1 on the preceding items and $T_{1n^*}(\theta)$ [except, of course, for the summed scores of 0 and n^* that involve only a single product].

For polytomous items, again excepting scores less than the number of response categories for the item and scores near the maximum attainable, for each value of n^* the sum involves k_{n^*} terms. For tests with more than the four score categories illustrated here, seven-point rectangular quadrature is not adequate. However, the relative robustness of the method to quadrature is illustrated by the fact that if quadrature in the example in Table 1 is increased from the seven points at unit intervals shown in the table to 46 points between −4.5 and 4.5 with an interval of .2, the final values of the proportion in each score group differ by less than .0001, and the values of the EAPs differ by less than .01.

Example Applications

Polytomous Data

Data from the North Carolina End-of-Grade 3 Social Studies exam were used. The test consisted of three open-ended items, which were administered to 23,374 students in the spring of 1993. The responses

Table 1
A Numerical Example for Three Binary Items for Values of i and j and $n^* = 0, 1, 2$ at Seven Levels of θ

Variable	i	j	θ = −3	θ = −2	θ = −1	θ = 0	θ = 1	θ = 2	θ = 3
Initialization: IRF Ordinates									
$T_{ji}(\theta)$	0	0	.73106	.62246	.50000	.37754	.26894	.18243	.11920
		1	.26894	.37754	.50000	.62246	.73106	.81757	.88080
	1	0	.95257	.88080	.73106	.50000	.26894	.11920	.04743
		1	.04743	.11920	.26894	.50000	.73106	.88080	.95257
	2	0	.99753	.98901	.95257	.81757	.50000	.18243	.04743
		1	.00247	.01099	.04743	.18243	.50000	.81757	.95257
Initialization: For $n^* = 0$									
$L_0^0(\theta) = T_{00}(\theta)$		0	.73106	.62246	.50000	.37754	.26894	.18243	.11920
$L_1^0(\theta) = T_{10}(\theta)$		1	.26894	.37754	.50000	.62246	.73106	.81757	.88080
For $n^* = 1$									
$L_0^1(\theta) = L_0^0(\theta)T_{01}(\theta)$		0	.69639	.54826	.36553	.18877	.07233	.02175	.00565
$L_1^1(\theta) = L_1^0(\theta)T_{01}(\theta) + L_0^0(\theta)T_{11}(\theta)$		1	.29086	.40674	.50000	.50000	.39322	.25814	.15532
$L_2^1(\theta) = L_1^0(\theta)T_{11}(\theta)$		2	.01276	.04500	.13447	.31123	.53445	.72012	.83902
For $n^* = 2$									
$L_0^2(\theta) = L_0^1(\theta)T_{02}(\theta)$		0	.69467	.54224	.34819	.15433	.03616	.00397	.00027
$L_1^2(\theta) = L_1^1(\theta)T_{02}(\theta) + L_0^1(\theta)T_{12}(\theta)$		1	.29186	.40829	.49362	.44322	.23278	.06487	.01275
$L_2^2(\theta) = L_1^2(\theta)T_{02}(\theta) + L_1^1(\theta)T_{12}(\theta)$		2	.01344	.04898	.15181	.34567	.46384	.34241	.18775
$L_3^2(\theta) = L_2^1(\theta)T_{12}(\theta)$		3	.00003	.00049	.00638	.05678	.26722	.58875	.79923

to the open-ended items were rated on a 4-point scale—item scores ranged from 0–3 and summed scores ranged from 0–9. The parameter estimates for these items were obtained using Samejima's (1969) graded model and the computer program MULTILOG (Thissen, 1991) (see Table 2).

Figure 1 shows the posterior density for the three response patterns that had a summed score of 1: 100, 010, and 001. Of these response patterns, 001 was the most frequently observed in the data (2,830 examinees), followed by 010 (2,008 examinees), and then 100 (811 examinees). This pattern reflected the differential difficulty of the items (see Table 2): Item 3 had the lowest threshold (b) for Category 1 ($b_1 = .08$), so it was most likely that if an examinee received a single 1 and two 0s, the 1 would be for Item 3. It was not much more difficult to obtain a score of 1 on Item 2 ($b_1 = .12$), but obtaining a 1 on Item 1 was substantially more difficult ($b_1 = .65$).

Table 2 shows that Item 2 was the most discriminating (a), followed by Item 1, and then Item 3. Thus, the posterior distributions for the three response patterns shown in Figure 1 had different locations: The averages, or EAPs, for the response patterns 010, 100, and 001 were .08, −.15, and −.38, respectively. Response-pattern scaled scores reflect the differences among examinees with different response patterns. For example, examinees with response pattern 010 had somewhat higher θs than those with response patterns 001 or 100.

Figure 1 also shows the posterior distribution for all examinees who obtained a summed score of 1, which is the total of the three posterior distributions for response patterns 100, 010, and 001, computed using the item parameters in Table 2 and the recursive algorithm described above. The average of this posterior distribution, or $\mathrm{EAP}(\theta|j = \Sigma k_i)$, may be used to describe the average ability of examinees who obtained a summed score of 1, in the same way that the three different response pattern EAPs described the

Table 2
Item Parameter Estimates for the Three-Item Social Studies Test [θ was Distributed N(0,1)]

Item	a	b_1	b_2	b_3
1	1.87	.65	1.97	3.14
2	2.66	.12	1.57	2.69
3	1.24	.08	2.03	4.30

ability of examinees with a particular response pattern. For the example in Figure 1, the summed-score EAP $\left[\mathrm{EAP}(\theta|j=1)\right]$ was −.18, and the associated SD was .61.

The summed-score EAP is, approximately, a weighted average of the EAPs for the response patterns that yield that summed score, and the SD of the summed-score EAP tends to be slightly larger than the SD of the EAPs for most of the patterns with that summed score. For the example of a summed score of 1 on this test, the SDs for the pattern EAPs were .60 (001), .54 (010), and .57 (100). Thus, although there was some loss of precision entailed in computing scaled scores for each summed-score group instead of for each response-pattern group, that loss of precision appears small. For the most frequent response pattern with a summed score of 1 (001), the difference between the SD of the pattern posterior and that for the summed score posterior was .01.

Figure 1
The Posterior Density for the Three Response Patterns on the Grade 3 Social Studies Test That Had a Summed Score of 1 (100, 010, and 001) and the Posterior Distribution for all Examinees Obtaining a Summed Score of 1 (θ was Standardized)

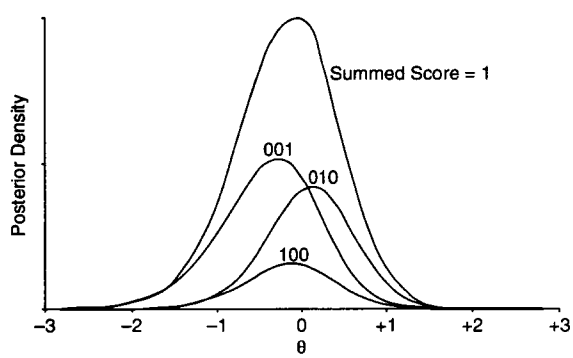


Table 3 shows the range of response pattern EAPs and the associated SDs for each summed score on this three-item test, as well as the EAPs and SDs for the most common and least common response patterns, with the summed-score EAPs and SDs. For items that differ in discrimination as these did, the response pattern EAPs may be highly variable (as much as a standard unit) within any particular summed score; however, the variation was mostly accounted for by the few examinees who produced unusual response patterns. Most of the responses were in a few common response patterns, and the summed-score EAPs and SDs were very similar to those for the most common response patterns within each score.

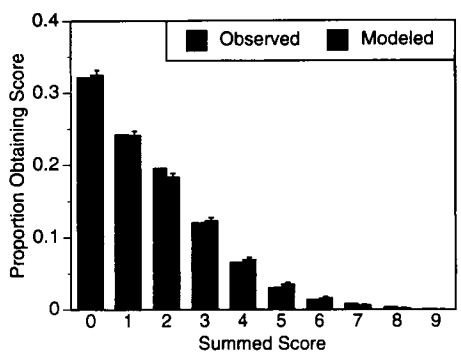
In general, the increase in the SDs from the smallest values for the response-pattern EAPs to the summed-score EAPs was approximately 10%. This is similar to the values that Birnbaum (1968, p. 477) reported in his study of the difference between summed scores and response-pattern scores, and is also approximately the same value observed in most applications of this procedure. The 10% loss of precision (on the scale of the SDs, which are reported as the standard errors of EAP scaled scores) represents the cost of assigning scores at the summed-score level rather than at the more precise response-pattern level.

Table 3
For Each Summed Score for the Three-Item Social Studies Test, Response-Pattern EAP Scaled Scores and Their SDs, Summed Score EAPs and Their SDs, and Observed and Modeled Proportion in Each Score Group

Summed Score	Pattern EAP Range	Pattern SD Range	Most Common Pattern EAP	Most Common Pattern SD	Least Common Pattern EAP	Least Common Pattern SD	(θ|Score) EAP	(θ|Score) SD	Observed Score Group Proportion	Modeled Score Group Proportion
0	−.88	.70					−.88	.70	.321	.325
1	−.38 − .08	.54 − .60	−.38	.60	−.15	.57	−.18	.61	.242	.241
2	−.21 − .59	.50 − .62	.39	.51	0.00	.61	.33	.57	.195	.183
3	−.18 − 1.10	.48 − .73	.80	.48	.74	.73	.74	.55	.120	.123
4	.44 − 1.45	.49 − .65	1.00	.49	.44	.47	1.12	.54	.065	.069
5	.76 − 1.87	.48 − .67	1.56	.48	.76	.65	1.48	.54	.030	.035
6	1.42 − 2.21	.47 − .63	1.88	.47	2.21	.63	1.84	.54	.014	.016
7	1.62 − 2.41	.49 − .60	2.36	.51	1.62	.60	2.21	.54	.008	.006
8	2.29 − 2.72	.53 − .54	2.72	.54	2.29	.53	2.62	.56	.003	.002
9	2.99	.56					2.99	.56	.0007	.0003

Table 3 also shows the observed proportions obtaining each summed score on this test and the modeled proportion computed using Equation 1; Figure 2 shows those two distributions. The modeled distribution is very close to the observed distribution. The distribution is very skewed, because this test was extraordinarily difficult. Nevertheless, the modeled proportions obtaining each score were computed using a Gaussian distribution as $\phi(\theta)$, illustrating Lord's (1953) argument that the summed-score distribution does not directly reflect the shape of the population distribution for the trait.

Figure 2
The Observed Proportions Obtaining Each Summed Score on the Grade 3 Social Studies Test and the Modeled Proportion Computed Using Equation 6 (Error Bars Show Pointwise Twice the Binomial Standard Error for the Proportions)



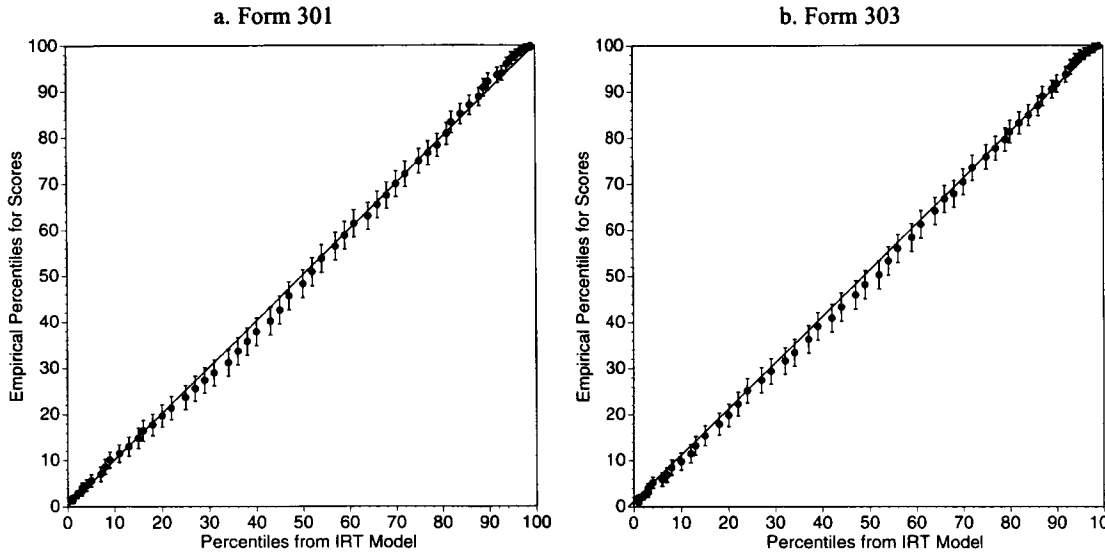
Dichotomous Data

For the second example, data from the spring 1992 administration of two preliminary forms of the North Carolina End-of-Grade 3 Mathematics exam were used. The two forms each contained 80 four-alternative multiple-choice items. Form 301 was administered to 1,053 examinees, and Form 303 was administered to 1,071 examinees. Three-parameter logistic item parameter estimates were computed using MULTILOG (Thissen, 1991), with the population distribution specified as N(0,1).

Figure 3 shows plots of the empirical percentiles for each observed summed score plotted against the

model-derived percentiles computed by accumulating P_j for Forms 301 (Figure 3a) and 303 (Figure 3b). The maximum absolute difference between the observed and model-based percentiles was less than 3.0 for both plots; that is approximately twice the maximum pointwise standard error of the empirical percentiles, which for N ≈ 1,000 was approximately 1.5 near the middle of the distribution. Because Figure 3 shows the data from which the item parameters were estimated, they show an aspect of the goodness of fit of the IRT model (including the underlying normal population distribution) to the data: The model reproduced the observed score distribution fairly accurately.

Figure 3
Empirical Percentiles for Each Observed Summed Score Plotted Against the Model-Derived Percentiles Computed by Accumulating P_j (Error Bars Show Pointwise Twice the Binomial Standard Error for the Percentiles)



Discussion

As Yen (1984) noted, IRT scaled scores can be effectively computed for each observed summed score on a test, providing the usefulness of the IRT score scale without the problems associated with response-pattern scoring. Although some loss of information follows from the simplification of scoring from response patterns to summed scores, that loss of information is small—the corresponding change in the reported standard error would often not result in a visible change in the number of decimals usually reported. The loss may be counterbalanced by more practical or socially-acceptable score reporting: Scaled score reporting based on summed scores is obviously more practical than response-pattern scoring, because the scaled scores may be obtained from a compact score-translation table; that is not possible, in general, for pattern scoring. Score reporting based on summed scores is often more socially acceptable for many consumers of test scores, because questions of perceived unfairness arise when examinees with the same number-correct score are given different scaled scores.

For the most part, the results reported here correspond to Yen's findings with modal scores based on the likelihood of the item responses alone, without consideration of the population distribution. The differences between the results found here and those reported by Yen (1984) may be accounted for by the fact

that the Gaussian population distribution and $\mathrm{EAP}(\theta|j=\Sigma k_i)$, were used here in place of the likelihood alone and its mode. Yen reported greatly increased variability with scores at the extremes of the distribution; those problems do not arise when the full model is used.

In addition, when the population distribution is included in the IRT model, computation of the observed score distribution itself is straightforward. The expected distribution can be used to provide smoothed percentile tables for the current form of the test, or preoperational percentile tables for tests assembled based on IRT, using the item parameters and the parameters of the population distribution.

If the population distribution assumed in the IRT model does not represent the distribution of θ well for the examinees, then the inferred score distribution will depart from the observed score distribution, which has both positive and negative implications. A positive implication is that such a departure should be useful as a diagnostic suggesting misspecification of the population distribution. On the negative side, the inferred score distribution will not be accurate as a source of preoperational percentile tables or for other similar uses. The extent to which the inferred score distribution might be sensitive to misspecification of the population distribution has not yet been examined. In all cases in which it has been used thus far with the North Carolina testing program, the assumption of a normal population distribution for θ has produced score distributions very much like the observed score distributions.

In principle, the recursive algorithm used here for the computation of $\mathrm{EAP}(\theta|j=\Sigma k_i)$, $\mathrm{SD}(\theta|j=\Sigma k_i)$, and P_j may be used for tests that combine binary-scored multiple-choice sections with open-ended items scored in multiple categories, simply by accumulating the "points" for each item into the score j. However, in practice this solution may not produce scaled scores with adequate precision. The problem is that rated "points" on open-ended items may reflect very different changes in scaled scores than do the "points" associated with each correct response to the multiple-choice items. Indeed, the rated "points" for the open-ended items may be associated with very different increases in scaled scores at different levels on the θ continuum.

For example, for the third grade Social Studies test (see Table 3), an increase of one "point" from 0 to 1 was associated with an increase in the scaled score from −.88 to −.18, or .7 standard units; however, an increase of one "point" between scores of 8 and 9 was associated with an increase in the scaled score of only half as much (.37). For the multiple-choice tests (percentiles are depicted in Figure 3), an increase in the summed score of one "point" was associated with a difference in scaled scores of between .05 and .2 standard units, depending on the location on the score scale. To some extent, this difference in the relative value of "points" may be adjusted by scoring the open-ended items with more "points;" this analysis indicates that the difference between open-ended rating values was approximately 4–5 multiple-choice points.

A better solution to the problem of combining binary-scored multiple-choice sections with open-ended items scored in multiple categories may involve a hybridization of summed-score and response-pattern computation of scaled scores. To implement this approach, compute $L_j^{\mathrm{MC}}(\theta)$ for the multiple-choice section and $L_{j'}^{\mathrm{OE}}(\theta)$ for the open-ended section using the recursive algorithm. Next, for each combination of a given summed score on the multiple-choice section with any summed score on the open-ended section, compute the product $L_j^{\mathrm{MC}}(\theta)L_{j'}^{\mathrm{OE}}(\theta)\phi(\theta)$. Then, taking these products as the posterior density for that response pattern (score j on the multiple-choice section and score j' on the open-ended section), compute the expected values and SDs for each of those posterior distributions. The resulting two-way score translation table would provide scaled scores and their standard errors for each "response pattern," where the "pattern" refers to the ordered pair (score j on the multiple-choice section, score j' on the open-ended section). This procedure would offer many of the practical advantages of summed scores, and it would preserve the differences in scaled scores that may be associated with very different values of "points" on the multiple-choice and open-ended sections.

References

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

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6,* 431–444.

Lord, F. M. (1953). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement, 13,* 517–548.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading MA: Addison-Wesley.

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

Masters, G. N., & Wright, B. D. (1984). The essential process in a family of measurement models. *Psychometrika, 49,* 529–544.

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika, 56,* 611–630.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research. Expanded edition, University of Chicago Press, 1980.

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

Stroud, A. H. (1974). *Numerical quadrature and solution of ordinary differential equations.* New York: Springer-Verlag.

Thissen, D. (1991). *MULTILOG user's guide: Multiple, categorical item analysis and test scoring using item response theory.* Chicago: Scientific Software.

Thissen, D., Pommerich, M., & Williams, V. S. L. (1993, June). *Some algorithms for computing E(θ\|summed score), and the implied score distribution, using item response theory.* Paper presented at the meeting of the Psychometric Society, Berkeley CA.

Tierney, L. (1990). *Lisp-Stat: An object-oriented environment for statistical computing and dynamic graphics.* New York: Wiley.

Walsh, J. E. (1963). Corrections to two papers concerned with binomial events. *Sankhyā, 25,* Series A, 427.

Wilson, D. T., Wood, R., & Gibbons, R. (1991). *Testfact: Test scoring, item statistics, and item factor analysis.* Chicago: Scientific Software.

Yen, W. M. (1984). Obtaining maximum likelihood trait estimates from number-correct scores for the three-parameter logistic model. *Journal of Educational Measurement, 21,* 93–111.

Acknowledgments

The research reported here was supported by the North Carolina Department of Public Instruction, in conjunction with the development of the North Carolina End-of-Grade Testing Program. The authors thank Richard Luecht, Robert McKinley, Robert Mislevy, James Ramsay, and Linda Wightman for their help in the course of this work.

Author's Address

Send requests for reprints or further information to David Thissen, L. L. Thurstone Psychometric Laboratory, University of North Carolina, CB #3270, Davie Hall, Chapel Hill NC 27599-3270, U.S.A.