Psychological Assessment
2000, Vol. 12, No. 3, 354–359

1040-3590/00/$5.00 DOI: 10.1037//1040-3590.12.3.354

Summed-Score Linking Using Item Response Theory: Application to Depression Measurement

Maria Orlando and Cathy D. Sherbourne
RAND

David Thissen
University of North Carolina at Chapel Hill

An item response theory (IRT) approach to test linking based on summed scores is presented and demonstrated by calibrating a modified 23-item version of the Center for Epidemiologic Studies Depression Scale (CES-D) to the standard 20-item CES-D. Data are from the Depression Patient Outcomes Research Team, II, which used a modified CES-D to measure risk for depression. Responses (N = 1,120) to items on both the original and modified versions were calibrated simultaneously using F. Samejima's (1969, 1997) graded IRT model. The 2 scales were linked on the basis of derived summed-score-to-IRT-score translation tables. The established cut score of 16 on the standard CES-D corresponded most closely to a summed score of 20 on the modified version. The IRT summed-score approach to test linking is a straightforward, valid, and practical method that can be applied in a variety of situations.

Item response theory (IRT) comprises a collection of modeling techniques for the analysis of items, tests, and people. When the assumptions of the IRT model are met, this collection of techniques offers many advantages over traditional test theory and can be a powerful tool for test construction (Embretson, 1996). Because developments in IRT have been driven primarily by the increasing importance of educational testing (Hambleton & Swaminathan, 1985; Lord, 1980; Wainer et al., 1990), much of the language used to describe the theory is specific to this context (e.g., proficiency or ability, number correct, examinee, test). However, investigators can apply IRT in any situation in which responses to questions are intended to relate to some unidimensional construct.[1]

IRT is said to have a "built-in" linking mechanism (Embretson, 1996). *Linking* is a general term used to describe the comparison of results from two or more separate assessments and can be used to refer to both equating and calibration. Once item parameters are estimated for a population with an IRT model, investigators may calculate comparable scores on a given construct for respondents from that population who did not answer the same questions, without intermediate equating steps. The requirements for equating are stringent, but calibrating two tests of different lengths for a specific purpose is less stringent and can easily be achieved using an IRT approach (Linn, 1993; Mislevy, 1992).

Maria Orlando and Cathy D. Sherbourne, RAND Health Division, Santa Monica, California; David Thissen, Department of Psychology, University of North Carolina at Chapel Hill.

This work was funded by Grant R01-HS08349 from the Agency for Healthcare Research and Quality (formerly the Agency for Health Care Policy and Research), Grant P50MH54623 from the National Institute of Mental Health, and Grant 96-42901A-HE from the John D. and Catherine T. MacArthur Foundation.

Correspondence concerning this article should be addressed to Maria Orlando, RAND, 1700 Main Street, Box 2138, Santa Monica, California 90407-2138. Electronic mail may be sent to Maria_Orlando@rand.org.

Although linking is fairly straightforward in IRT applications, a practical problem is that scores are linked on the latent θ continuum, or IRT-score scale. In general, the observed summed-score scale is more intuitively understood, and the ultimate goal of many linking tasks is to calibrate scores from different tests on this scale. Regardless of the method chosen, calibration in this context entails finding the summed score on one test form that corresponds to the summed score on a second test form. Until recently, IRT methods have not provided a direct answer to this question. With the exception of the Rasch (1960) family of models, in which the summed score is a sufficient statistic for the IRT score, there has been no one-to-one transformation from IRT score to summed score.

Lord and Wingersky (1984) introduced a method to calculate the IRT likelihood for each summed score on a test. This method was recently used as the basis for the computation of IRT-scale scores for each summed score (Thissen, Pommerich, Billeaud, & Williams, 1995). Zeng and Kolen (1995) integrated this method into an IRT observed-score equating application for tests with dichotomous items. Despite the advantages of working with summed scores rather than IRT scores, investigators have not used this approach extensively. In this article, we describe the calculation of IRT scores for each summed score on a test and illustrate its use by linking two depression scales—the traditional 20-item Center for Epidemiologic Studies Depression Scale (CES-D; Radloff, 1977) and a 23-item variant (Daniel Ford, personal communication, August 1995)—with data from a longitudinal survey.

Background

The item characteristic curve, or trace line, is at the core of IRT and is most commonly defined as a logistic function. One result of

[1] IRT has been extended to the multidimensional case (Reckase, 1997), but in this article, we refer only to unidimensional IRT applications.

an IRT calibration is a set of continuous trace lines that describe the probability of endorsing each response category of an item given the scale value of the construct being measured. For dichotomous items, the probability of a negative response is high for low values of the construct being measured (denoted as θ), decreasing as θ increases. Conversely, the probability of a positive response is low at the low end of the θ continuum and increases to 1 as θ increases. The trace lines from an IRT calibration are continuous functions that reflect this pattern of responding.

Estimating a score using IRT methods involves the multiplication of all the trace lines corresponding to a person's responses to each question. For example, on a three-item test, a response of *no* to the first two items and *yes* to the third item would result in three trace lines similar to those shown in the top panel of Figure 1. A posterior distribution is formed by multiplying the three trace lines together with a prior distribution (that is usually normal); this product appears in the bottom panel of Figure 1, and the IRT score calculates as the average (or *expected a posteriori,* or *EAP*) of this posterior distribution.

Assuming a set of items with known parameters related to a construct, one can calculate an IRT score in this way for any given response pattern. These IRT scores are considered to be scale-equivalent and comparable regardless of the set of items presented (Lord, 1980). In most applications of IRT, however, investigators do not know the item parameters prior to administration; they are estimated simultaneously with the IRT scores. In this context, the linking of two or more test forms is still fairly straightforward, provided there are some overlapping items on the forms. Investigators can estimate item parameters on the basis of responses to the items from both tests, unique and overlapping, as if they comprised a single test, and IRT scores can be derived on the basis of these item parameters for the response patterns in the data for each test.

Method

The goal of many linking tasks is to find the summed score on one form of the test that corresponds to a given summed score on the other form of the test, or to compare the mean score of one administration or cohort to another in terms of number correct. These outcomes require that results be applicable to the summed-score scale on the test. Applying the summed-score-to-IRT-score approach presented here takes advantage of the straightforward linking process built in to IRT methodology as well as the utility and practicality of linking tests on the summed-score scale.

Obtaining an IRT score corresponding to a summed score, rather than to a particular pattern of responses, requires finding the average of a posterior distribution similar to that which appears in the bottom panel of Figure 1. The difference is that the desired posterior is the sum of the posteriors for all of the possible response patterns that yield that summed score. Because the number of response patterns that yield a particular summed score may be large, brute force computation of the sum of their posteriors is impractical. Lord and Wingersky (1984) briefly described a method of computing

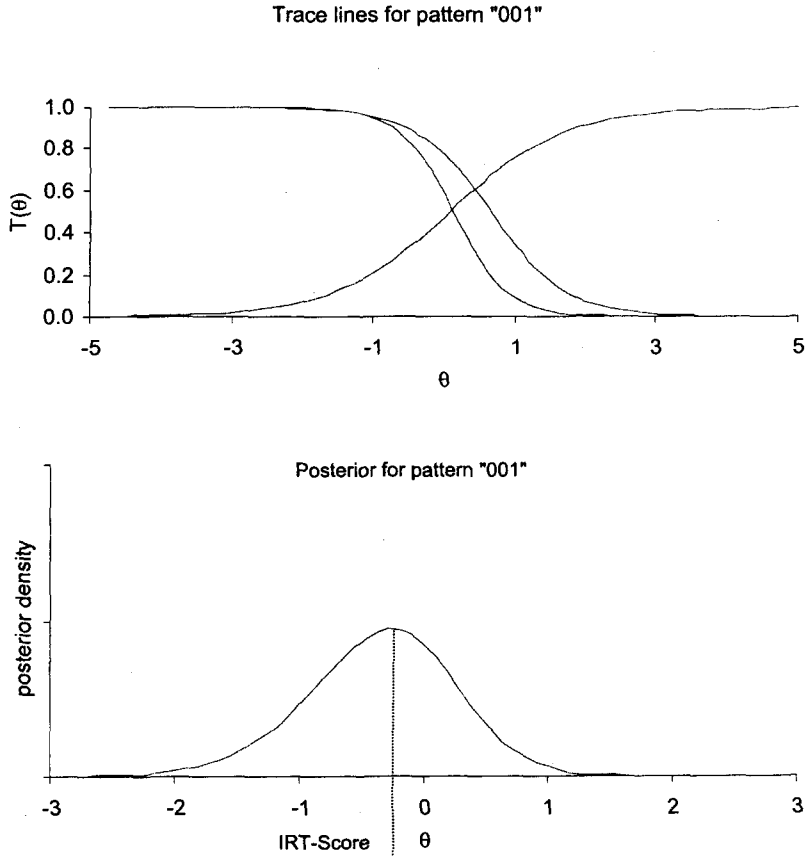

Figure 1. Top panel: trace lines for responding *no* to two items and *yes* to one item on a three-item test. Bottom panel: posterior for this response pattern. IRT = item response theory.

the IRT model-predicted distributions for each summed score, with Thissen et al. (1995) going into more detail. This approach uses a recursive algorithm that builds the joint likelihood for each score group item by item. Using this algorithm, the likelihoods for summed scores 0 through k (the number of response categories) begin as the trace line for a response in the corresponding category on the first item. Then, each item is included sequentially, with trace lines for each response category being multiplied into the appropriate summed-score likelihood. When one has considered all the items, this algorithm has effectively collected all the pattern likelihoods corresponding to each summed score to form a joint likelihood. Then, one can calculate the average (or *EAP*) value as the IRT score associated with that summed score. (Please see the Appendix for details.)

To link two or more tests on the summed-score scale using IRT (assuming item parameters are not already available), all test items—those unique to one test and those appearing on two or more tests—are first calibrated simultaneously to obtain a common set of item parameters. These item parameters are then grouped according to test form so that each form has a set of parameters for all of its items (unique and overlapping). The recursive algorithm is then applied to each test's set of parameters to get the average IRT score associated with each possible summed score on the test. The resultant summed-score-to-IRT-score translation tables for each test are then used to link the tests. The example that follows uses this methodology to find the score on a 23-item variant of the CES-D that corresponds to a score of 16 on the traditional 20-item CES-D.

Example

Data for this example come from the Depression Patient Outcomes Research Team, II (PORT), a longitudinal study of quality improvement for depression in managed care primary care practices. Details of the study design are available elsewhere (Wells, 1999). We administered a variant of the CES-D, developed to incorporate *Diagnostic and Statistical Manual of Mental Disorders* (4th ed.; *DSM–IV*; American Psychiatric Association, 1994) criteria for depression, to assess risk for depression in selected waves of this study. This scale contained 13 items from the CES-D, a widely used 20-item depression summary scale, and 10 new items. In a subsequent wave of the study (with an 83% retention rate from baseline), we administered a 30-item scale to 1,120 participants. This scale contained the 20 CES-D items and the 10 new items used in the 23-item scale. Figure 2 depicts the relationships among these three sets of items.

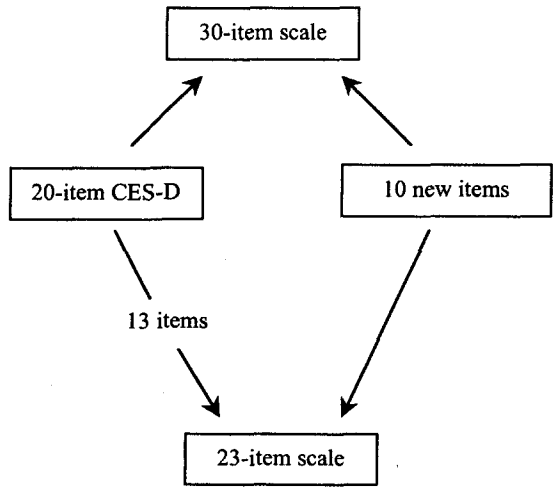

Figure 2. Relationships among the 30 calibrated depression items. CES-D = Center for Epidemiologic Studies Depression Scale.

Responses to items on the CES-D, the 23-item scale, and the 30-item scale all use a four-category, Likert-type response scale in which participants are asked to indicate the extent to which they have experienced the feeling or condition expressed in the item stem during the past week. Response categories are *rarely or none, some, occasional,* and *most or all.* Some items are reverse-scored, so that for all items, a higher score reflects more depression. Possible total scores on the 20-item scale range from 0–60 and on the 23-item scale from 0–69.

The CES-D has an established cut-score total score of 16, indicating risk for depression (Mulrow et al., 1995; Zich, Attkisson, & Greenfield, 1990). We used the IRT summed-score methodology to identify an equivalent cut score on the 23-item scale.

Before linking the two scales, we established their similarity with a series of descriptive analyses. First, we calculated the correlation between the 23-item scale and two subscales of the 36-Item Short-Form Health Survey (SF-36; Ware & Sherbourne, 1992), the Physical Component Summary Scale 12 (PCS-12) and the Mental Component Summary Scale 12 (MCS-12) and compared them to the corresponding correlations for the traditional 20-item CES-D. These relationships were very similar ($r = -0.35$, $r = -0.77$ for the 23-item scale; $r = -0.30$, $r = -0.77$ for the 20-item scale). We also calculated the correlation between the 20-item scale and the 23-item scale at 18 months, as well as the correlation between the non-overlapping items on the two scales (7 items from the 20-item scale and 13 items from the 23-item scale). The relationship between the two scales was very strong ($r = .97$), as would be expected given the overlap of items. In addition, the correlation between the non-overlapping item sets was also substantial ($r = .79$). This empirical evidence suggests that the 23-item scale measures depression similarly to the 20 items from the CES-D, and that it is reasonable to calculate equivalent cut scores on the two scales.

Next, we performed a principal components analysis to establish the unidimensionality of the 30 items. Although there were five eigenvalues greater than 1, the first eigenvalue (13.4) was substantially greater than the next four (1.6, 1.5, 1.4, 1.1). In addition, 29 of the 30 items had standardized factor loadings greater than .35, ranging from .28 to .81, with an average of .65. These results indicate that the factor structure of the 30 items is sufficiently unidimensional for application of IRT.

As a first step in the linking process, we used MULTILOG (Thissen, 1991) to simultaneously calibrate all 30 depression items on the basis of data from respondents who answered all 30 questions. Assuming the 30 items are sufficiently unidimensional, this calibrates any shorter test assembled from these 30 items. Because of the ordered nature of the CES-D item responses, we used Samejima's (1969, 1997) graded IRT model to fit these data. The parameterization of this model implies that the probability of endorsing the lowest response category is high at the low end of the θ continuum and decreases to 0; that the probability of endorsing the highest response category begins at 0 and increases to 1 with θ; and that the probability of endorsing the intermediate response categories rises and falls sequentially with increasing θ. The model converged nicely to a solution, yielding a coherent set of item parameters.

Table 1
IRT-Score-to-Summed-Score Conversion Tables for the 20- and 23-Item Scales

20-item scale			23-item scale		
Summed score	IRT score	*SD*	Summed score	IRT score	*SD*
0	−2.1	0.52	0	−2.2	0.52
—	—		—	—	
15	−0.2	0.025	15	−0.41	0.023
16	−0.14	0.025	16	−0.35	0.023
17	−0.08	0.024	17	−0.29	0.023
18	0	0.024	18	−0.23	0.022
19	0.037	0.023	19	−0.17	0.022
20	0.093	0.022	20	−0.12	0.022
21	0.15	0.022	21	−0.064	0.021
22	0.2	0.021	22	−0.012	0.021
23	0.26	0.021	23	0.04	0.021
24	0.31	0.020	24	0.091	0.020
25	0.36	0.019	25	0.14	0.020
—	—		—	—	
60	3	0.43	69	3.2	0.43

Note. The 20-item scale is the original Center for Epidemiologic Studies Depression Scale (CES-D). The 23-item scale contains 13 items from the original CES-D and 10 items designed to assess *DSM–IV* depression elements. IRT = item response theory.

We used the item parameters from the 20-item scale and item parameters from the 23-item scale as separate input into the program SS_IRT[2] to estimate the IRT score corresponding to each summed score for each scale. Table 1 displays a segment of the summed-score-to-IRT-score translation tables for these two scales. On the 20-item scale, a summed score of 16 corresponds to an IRT score of −0.14. This means that the cut score of 16 distinguishes between people above an IRT score of −0.14 and below an IRT score of −0.14. To find an equivalent cut score on the 23-item scale, we identified the IRT score in the 23-item scale translation table that is closest to −0.14, the IRT score of −0.12. On the 23-item scale, the summed score that corresponds to an IRT score of −0.12 is 20.

In order to examine the validity of the cut score of 20, we compared the rates of respondents at the 18-month wave classified as depressed using both the 20 CES-D items (and corresponding cut score of 16) and the 23-item scale (with its cut score of 20). Table 2 is a weighted cross-classification table for these two criteria. Nearly 95% of the sample are classified in the same way regardless of the criterion used.

Table 2
Weighted Percent of Respondents Classified as Depressed According to the 20-Item Cut Point of 16 and the 23-Item Cut Point of 20

	20-item scale		
23-item scale	Below 16	At or above 16	
Below 20	41.34	3.83	45.16
At or above 20	2.73	52.11	54.84
	44.07	55.93	

Note. The 20-item scale is the original Center for Epidemiologic Studies Depression Scale (CES-D). The 23-item scale contains 13 items from the original CES-D and 10 items designed to assess *DSM–IV* depression elements.

Table 3
Weighted Prevalence and Concordance Rates for Respondents Above Cut Point 20 on the 23-Item Scale and Respondents With CIDI Depression Diagnosis at Baseline and 24 Months

	Prevalence		Concordance rates			
Time point	% above 23-item cut point	% with positive CIDI diagnosis	Sensitivity	Specificity	PPV	NPV
Baseline	82.2	55.9	88.3	25.5	60.1	63.3
24 months	68.2	34.6	85.8	41.2	43.5	84.6

Note. CIDI = Composite International Diagnostic Interview; PPV = positive predictive value; NPV = negative predictive value.

In addition to the general survey instrument, study participants at baseline and at 24 months completed the Composite International Diagnostic Interview ([CIDI] 1990), a full diagnostic interview. As another check on the validity of the cut score and to provide a comparison to published reports of the 20-item CES-D, we used the 12-month depression diagnosis indicator derived from the CIDI as the standard to evaluate the sensitivity (the probability of screening positive given that the diagnosis is present), specificity (the probability of screening negative given that the diagnosis is absent), and positive/negative predictive values (the probability that the diagnosis is actually present/absent given that the screen was positive/negative) of the 23-item depression scale cut score of 20 at baseline and at 24 months. These weighted prevalence estimates and concordance rates are reported in Table 3.

Although the specificity of the scale is low, indicating a high number of false positives, the sensitivity and predictive values at both time points are reasonable. The sensitivity of the 23-item scale is slightly higher and the specificity is lower than the summary measures of the CES-D reported by Mulrow et al. (1995), but all four concordance rates are comparable to those reported by Santor and Coyne (1997) for the traditional CES-D as well as several shortened versions of it in a primary care setting. In studies such as the PORT, it is desirable to minimize the number of true positive cases missed in screening; thus, it is most important that the sensitivity of the screener be high, even at the cost of potentially low specificity. In addition, the sensitivity may be particularly low in this example due to the nature of the PORT sampling design. Participants all passed an initial screening for probable symptoms before being administered the CIDI and the 23-item CES-D, so all participants either had a diagnosis of depression or had current depressive symptoms.

Conclusion

This article presents and illustrates a method of calibrating tests on the summed-score scale using an IRT approach. In the example

[2] The software SS_IRT is available by E-mail from Maria Orlando (Maria_Orlando@rand.org).

presented here, we identified the cut score on a 23-item variant of the CES-D that corresponds to the cut score of 16 on the traditional CES-D. The cut score for the 23-item scale identified respondents as being at risk for depression similarly to the cut score for the 20-item scale. In addition, the cut score of 20 demonstrated acceptable concordance rates with the CIDI at two time points.

The IRT summed-score approach to test linking is a straightforward, valid, and practical method that can be applied in a variety of situations. Questionnaires of various lengths, consisting of dichotomous, Likert-type, or combinations of response formats can be linked using this IRT approach, provided they are measuring the same construct and there is some degree of item overlap.

References

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

Composite International Diagnostic Interview: Version 1.0. (1990). Switzerland: World Health Organization.

Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment, 8,* 341–349.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston: Kluwer-Nijhoff.

Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education, 6,* 83–102.

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

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

Mislevy, R. J. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects.* Princeton, NJ: Educational Testing Service.

Mulrow, C. D., Williams, J. W., Gerety, M. B., Ramirez, G., Montiel, O. M., & Kerber, K. (1995). Case-finding instruments for depression in primary care settings. *Annals of Internal Medicine, 122,* 913–921.

Radloff, L. S. (1977). The CES-D Scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement, 1,* 385–401.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen, Denmark: Danmarks Paedagogiske Institut.

Reckase, M. D. (1997). A linear logistic multidimensional model for dichotomous item response data. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 271–286). New York: Springer-Verlag.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph, 34*(4, Pt. 2).

Samejima, F. (1997). Graded response model. In W. J. van der Linden and R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85–100). New York: Springer-Verlag.

Santor, D. A., & Coyne, J. C. (1997). Shortening the CES-D to improve its ability to detect cases of depression. *Psychological Assessment, 9,* 233–243.

Santor, D. A., & Ramsay, J. O. (1998). Progress in the technology of measurement: Applications of item response models. *Psychological Assessment, 10,* 345–359.

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

Thissen, D., & Orlando, M. (in press). Item response theory for items scored in two categories. In D. Thissen & H. Wainer (Eds.), *Test scoring.* Hillsdale, NJ: Erlbaum.

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

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, R. J., Steinberg, L., & Thissen, D. (1990). *Computerized adaptive testing: A primer.* Hillsdale, NJ: Earlbaum.

Ware, J. E., Jr., & Sherbourne, C. D. (1992). The MOS 36-Item Short-Form Health Survey: I. Conceptual framework and item selection. *Medical Care, 30,* 473–483.

Wells, K. B. (1999). The design of partners in care: Evaluating the cost-effectiveness of improving care for depression in primary care. *Social Psychiatry and Psychiatric Epidemiology, 34,* 20–29.

Zeng, L., & Kolen, M. J. (1995). An alternative approach for IRT observed-score equating of number-correct scores. *Applied Psychological Measurement, 19,* 231–240.

Zich, J. M., Attkisson, C. C., & Greenfield, T. K. (1990). Screening for depression in primary care clinics: The CES-D and the BDI. *International Journal of Psychiatry in Medicine, 20,* 259–277.

Appendix

IRT-Score to Summed-Score Translation

The following series of equations describes the construction of an IRT-score-to-summed-score translation table similar to Table 1 in the text. Because of its complexity, the process is explicated here for tests with dichotomous items. However, this type of conversion table can also be derived for tests with polytomous items, as in the example in this article, as well as for tests with a combination of dichotomous and polytomous items. Equations A1–A5 describe the calculation of the $EAP(\theta)$ and $SD(\theta)$, and Equations A6–A10 represent the recursive algorithm that is used to estimate the integrand in Equation A3. The values computed using Equation A4 may be tabulated and used as the IRT scaled-score transformation of the raw scores, and the values of Equation A5 may be used as a standard description of the uncertainty associated with those IRT scores.

For any IRT model for items indexed by i with item scores $u = 0, 1$, the likelihood for any summed score $x = \Sigma\, u_i$ is

$$L_x(\theta) = \sum_{(u_i)=x} L(\mathbf{u}|\theta),$$

where the summation is over all the response patterns that contain x correct responses. Note that each of the response pattern likelihoods, $L(u|\theta)$, and the sum, $L_x(\theta)$, are functions of θ—that is, this represents a summation of *curves,* each of which is the likelihood of a particular response pattern.

The joint likelihood for each response pattern is

$$L(\mathbf{u}|\theta) = \prod_i T_{u_i}(\theta)\phi(\theta), \tag{A1}$$

where $T_{u_i}(\theta)$ is the trace line for response u to item i, and $\phi(\theta)$ is the population density; a small example is shown in Figure 1.

Thus, the likelihood for each score is

$$L_x(\theta) = \sum_{(u_i)=x} \prod_i T_{u_i}(\theta)\phi(\theta), \tag{A2}$$

and so the probability of each score x is

$$P_x = \int L_x(\theta)d\theta \tag{A3}$$

(that is, the area under the curve $L_x(\theta)$ represents the modeled probability of each summed score just as the area under the curve $L(u|\theta)$ represents the probability of each response pattern).

Lord and Wingersky (1984) presented a recursive algorithm to compute the integrand in Equation A3, and Thissen et al. (1995) showed that, using this algorithm, it is straightforward to compute the average value of θ associated with each score,

$$EAP[\theta|x = \textstyle\sum \tau u_i] = \dfrac{\int \theta L_x(\theta)d\theta}{P_x}, \tag{A4}$$

and the corresponding standard deviation,

$$SD[\theta|x = \textstyle\sum u_i] = \left(\dfrac{\int [\theta - EAP(\theta|x = \sum u_i)]^2 L_x(\theta)d\theta}{P_x}\right)^{1/2}. \tag{A5}$$

The recursive algorithm to compute the integrand in Equation A3 follows, using the notation $i = 0, 1, \ldots I$ for the items and $x = 0, 1, \ldots X$ for the summed scores. In addition, $T_i(\theta)$ is the trace line for a positive response to item i, $L_x^*(\theta)$ is the interim value for the likelihood for summed score x, and $L_x(\theta)$ is the likelihood for summed score x for a set of items; the population distribution is $\phi(\theta)$. Begin by setting the likelihood for summed scores 0 and 1 equal to the trace lines correct and incorrect for the first item

$$L_0^*(\theta) = 1 - T_1(\theta), \tag{A6}$$

$$L_1^*(\theta) = T_1(\theta). \tag{A7}$$

Then, add each item i to the test, computing

$$L_0(\theta) = (1 - T_i(\theta))L_0(\theta), \tag{A8}$$

$$L_x(\theta) = T_i(\theta)L_{x-1}^*(\theta) + (1 - T_i(\theta))L_x^*(\theta), \tag{A9}$$

for $x = 1, \ldots i - 1$, and

$$L_i(\theta) = T_i(\theta)L_{i-1}^*(\theta); \tag{A10}$$

after each item is added, the new $L_x(\theta)$ replaces the previous $L_x^*(\theta)$ for all scores computed for the previous item.

The algorithm itself is completely general, requiring no particular parametric form, so it could be used with, for example, the kernel-smooth trace lines described by Santor and Ramsay (1998). However, the implementation of the algorithm used here assumes that dichotomous items are calibrated with either the one-, two-, or three-parameter logistic models and that Likert-type items use the graded logistic model. (This appendix follows the development by Thissen and Orlando [in press]; Thissen et al. [1995] offer a somewhat more terse presentation that generalizes the algorithm to the polytomous case.)

Received December 14, 1999
Revision received May 4, 2000
Accepted May 12, 2000 ■