Springer Proceedings in Mathematics & Statistics

L. Andries van der Ark
Daniel M. Bolt
Wen-Chung Wang
Jeffrey A. Douglas
Sy-Miin Chow *Editors*

Quantitative Psychology Research

The 79th Annual Meeting of the Psychometric Society, Madison, Wisconsin, 2014

Springer

Springer Proceedings in Mathematics & Statistics

Volume 140

More information about this series at http://www.springer.com/series/10533

Springer Proceedings in Mathematics & Statistics

This book series features volumes composed of select contributions from workshops and conferences in all areas of current research in mathematics and statistics, including operation research and optimization. In addition to an overall evaluation of the interest, scientific quality, and timeliness of each proposal at the hands of the publisher, individual contributions are all refereed to the high quality standards of leading journals in the field. Thus, this series provides the research community with well-edited, authoritative reports on developments in the most exciting areas of mathematical and statistical research today.

L. Andries van der Ark • Daniel M. Bolt
Wen-Chung Wang • Jeffrey A. Douglas
Sy-Miin Chow
Editors

Quantitative Psychology Research

The 79th Annual Meeting of the Psychometric Society, Madison, Wisconsin, 2014

Springer

Editors

L. Andries van der Ark
University of Amsterdam
Amsterdam, The Netherlands

Daniel M. Bolt
University of Wisconsin
Madison, WI, USA

Wen-Chung Wang
The Hong Kong Institute of Education
Hong Kong, Hong Kong SAR

Jeffrey A. Douglas
University of Illinois
Champaign, IL, USA

Sy-Miin Chow
The Penn State University
University Park, PA, USA

ISSN 2194-1009 ISSN 2194-1017 (electronic)
Springer Proceedings in Mathematics & Statistics
ISBN 978-3-319-19976-4 ISBN 978-3-319-19977-1 (eBook)
DOI 10.1007/978-3-319-19977-1

Library of Congress Control Number: 2015945591

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

Preface

This volume is a collection of presentations given at the 79th annual International Meeting of the Psychometric Society (IMPS) held at the University of Wisconsin-Madison, in Madison, Wisconsin during July 21–25, 2014. The meeting attracted 380 participants from 26 countries, with 242 papers being presented, along with 56 poster presentations, 5 pre-conference workshops, 5 keynote presentations, 4 invited speaker presentations, 4 state-of-the-art lectures, and 8 invited symposia. We thank the University of Wisconsin-Extension staff, as well as the faculty and students from the Department of Educational Psychology at the University of Wisconsin, Madison for hosting this very successful conference.

This volume continues a tradition started after the 77th meeting in Lincoln, Nebraska, of publishing a proceedings volume from the conference so as to allow presenters to quickly disseminate their ideas to the broader research community, while still undergoing a thorough review process. The 78th meeting in Arnhem was also followed by a proceedings. With the third proceedings, we now have a series that is expected to be continued next year with submissions from the 80th IMPS meeting in Beijing, China.

We asked the authors to use their presentations at the meeting as the basis of their chapters, possibly extended with new ideas or additional information. The result is a selection of 26 state-of-the-art chapters addressing a diverse set of topics, including item response theory, factor analysis, structural equation modelling, time series analysis, mediation analysis, propensity score methods, cognitive diagnostic models, and multi-level models, among others.

The proceedings of the 77th and 78th meeting were initiated by Roger E. Millsap, the editor of Psychometrika. Just before finalizing the proceedings of the 78th meeting, on May 9, 2014, Roger suddenly passed away. This volume is the first proceedings not initiated by Roger. We dedicate it to him.

Amsterdam, The Netherlands	L. Andries van der Ark
Madison, WI, USA	Daniel M. Bolt
Hong Kong, Hong Kong SAR	Wen-Chung Wang
Urbana-Champaign, IL, USA	Jeffrey A. Douglas
University Park, PA, USA	Sy-Miin Chow

Contents

Chapter 1
Extending the Use of Multidimensional IRT Calibration as Projection: Many-to-One Linking and Linear Computation of Projected Scores

David Thissen, Yang Liu, Brooke Magnus, and Hally Quinn

Abstract Two methods to make inferences about scores that would have been obtained on one test using responses obtained with a different test are *scale aligning* and *projection*. If both tests measure the same construct, scale aligning may be accomplished using the results of simultaneous calibration of the items from both tests with a unidimensional IRT model. If the tests measure distinct but related constructs, an alternative is the use of regression to predict scores on one test from scores on the other; when the score distribution is predicted, this is projection. Calibrated projection combines those two methods, using a multidimensional IRT (MIRT) model to simultaneously calibrate the items comprising two tests onto scales representing distinct constructs, and estimating the parameters describing the relation between the two scales. Then projection is done within the MIRT model. This presentation describes two extensions of calibrated projection: (1) the use of linear models to compute the projected scores and their error variances, and (2) projection from more than one test to a single test. The procedures are illustrated using data obtained with scales measuring closely related quality of life constructs.

Keywords linking • projection • calibration • scale aligning

1.1 Introduction

It is often desirable to obtain scores that are in some sense comparable from disparate tests that measure the same or closely related constructs. For example, the empirical examples in this presentation are motivated by the possibility that PROMIS® pediatric and adult scales may be used in the same cross-sectional or

D. Thissen (✉) • Y. Liu • B. Magnus • H. Quinn
Department of Psychology, The University of North Carolina
at Chapel Hill, Chapel Hill, NC 27599, USA
e-mail: dthissen@email.unc.edu; liuy0811@live.unc.edu;
brooke.magnus@unc.edu; hallyq@live.unc.edu

L.A. van der Ark et al. (eds.), *Quantitative Psychology Research*, Springer
Proceedings in Mathematics & Statistics 140, DOI 10.1007/978-3-319-19977-1_1

longitudinal research, with the (different) pediatric and adult scales each used within their age-range. Test score linking facilitates data analysis in such situations.

Holland (2007) provided a modern framework for test score linking; he wrote that "linking refers to the general class of transformations between the scores from one test and those of another, . . . linking methods can be divided into three basic categories called predicting, scale aligning, and equating." For linking scores from disparate scales, such as the PROMIS® pediatric and adult scales, only predicting scores from one with the other, or aligning the two scales, are viable candidates.

A commonly used method of scale aligning has been calibration, which uses item response theory (IRT) models and methods to place the items from each of two scales on the same metric. After that is done, standard computation of IRT scale scores from any subset of the items (which could include all of the items on only one scale) yields comparable scores. However, calibration has heretofore been limited to situations in which a unidimensional IRT model is suitable for all items from both scales jointly—that is, both scales measure the same construct.

For two scales that measure different constructs, even if the two constructs are highly related, predicting scores on one scale from those on the other yields more correct results. Such predictions are based on regression models, but often the regression model is elaborated to produce a distribution across the score range as a prediction; that is called projection.

Usually projection has been based on standard regression models, which consider the values of the predictor variable(s) fixed. *Calibrated projection* (Thissen et al. 2011) is a relatively new statistical procedure that uses IRT to link two measures, without considering the scores on the predictor scale to be fixed, and without the demand of conventional calibration that the two are measures of the same construct. In calibrated projection, a multidimensional IRT (MIRT) model is fitted to the item responses from the two measures: θ_1 represents the underlying construct measured by the first scale, with estimated slopes a_1 for each of the first scale's items and fixed values of 0.0 for the items of the second scale. θ_2 represents the underlying construct measured by the second scale, with estimated slopes a_2 for each of the second scale's items and fixed values of 0.0 for the items of the first scale. The correlation between θ_1 and θ_2 is estimated.

After calibration, the MIRT model may be used to provide IRT scale score estimates on the scale of the second measure, using only the item responses from the first measure. Figure 1.1 illustrates calibrated projection: The x-axis variable is θ_1, the underlying construct measured by the first scale (for Fig. 1.1, that is the PROMIS pediatric Anxiety scale), and the y-axis variable is θ_2, the underlying construct measured by the second scale (in Fig. 1.1, PROMIS adult Anxiety). The two latent variables are highly correlated, as indicated by the density ellipses around the regression line. Given the item responses on the pediatric Anxiety scale, IRT methods may be used to compute the implied distribution on θ_1; two of those are shown along the x-axis in Fig. 1.1, for summed scores of 13 and 44. The estimated relation between θ_1 and θ_2 is then used to project those distributions onto the y-axis, to yield the implied distributions on θ_2, the adult construct.

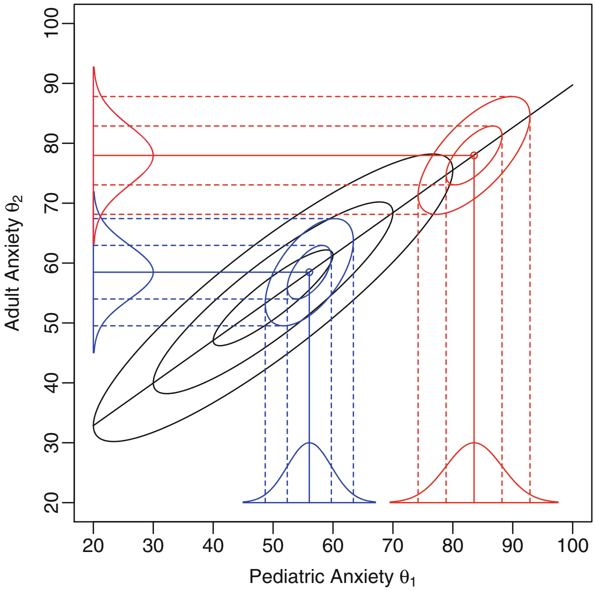

Fig. 1.1 The x-axis variable is θ_1, the underlying construct measured by the first scale, in this case the PROMIS pediatric Anxiety scale; the y-axis variable is θ_2, the underlying construct measured by the second scale, in this case the PROMIS adult Anxiety scale. Both scales report scores in T-score units. The correlation between the two latent variables is indicated by the large density ellipses. The implied distributions on θ_1 for summed scores of 13 (*blue*) and 44 (*red*) on the pediatric Anxiety scale are shown along the x-axis, along with the corresponding implied bivariate distributions, and those on θ_2, the adult Anxiety construct, along the y-axis

The means of the implied distributions on the θ dimensions are the IRT-based scale scores, and the standard deviations of those distributions are reported as the standard errors of those scores. The projection links the scales in the sense that each score on the pediatric scale yields a score on the adult metric.

In subsequent sections, we will illustrate calibrated projection from the PROMIS pediatric Anxiety (Irwin et al. 2010) scale to the corresponding adult scale (Pilkonis et al. 2011) and vice versa, using new data and pre-existing item parameters for the two PROMIS scales as the mechanism to link the results back to the original scales. Then we will describe a linear approximation to the IRT computations, and illustrate the extension of calibrated projection to use more than one scale as the basis for projection.

1.2 Calibrated Projection, Illustrated with PROMIS Anxiety

The original development of calibrated projection (Thissen et al. 2011) made use of the same data that were used to set the scale for the PROMIS Asthma Impact

Scale (PAIS), so there was no need to link a new set of data back to any existing scale. In contrast, the illustrations here are drawn from the linkage of the PROMIS pediatric and adult scales that measure similar constructs; both the pediatric and adult scales are now based on published item banks with reference metrics derived from their (separate) original calibrations. New data were collected for this project, from a sample of 874 persons in the age range 14–20 who responded to short forms of both the pediatric and adult PROMIS scales.

The item banks of all of the PROMIS scales comprise items with five response alternatives. The items have been calibrated using the graded IRT model (Samejima 1969, 1997), which describes the probability of each item response as a function of a set of item parameters (as and cs), and $\boldsymbol{\theta}$, the latent variable(s) measured by the scale, as follows: The conditional probability of response $u = 0, 1, \ldots, m - 1$ is

$$T_u(\boldsymbol{\theta}) = T_u^*(\boldsymbol{\theta}) - T_{u+1}^*(\boldsymbol{\theta}) \tag{1.1}$$

in which $T_u^*(\boldsymbol{\theta})$ is a curve tracing the probability of a response in category u or higher: $T_0^*(\boldsymbol{\theta}) = 1$, $T_m^*(\boldsymbol{\theta}) = 0$, and for $u = 1, 2, \ldots, m - 1$

$$T_u^*(\boldsymbol{\theta}) = \frac{1}{1 + \exp(-(\mathbf{a}'\boldsymbol{\theta} + c_u))}. \tag{1.2}$$

The original unidimensional parameters for the short-form items for the PROMIS pediatric and adult Anxiety scales are in Table 1.1, recast in a two-dimensional format in which θ_1 is the underlying construct measured by the pediatric Anxiety scale and θ_2 is the underlying construct measured by the adult Anxiety scale.

To begin the process of linking the two Anxiety scales with each other, and back to their original (published) scales, the item parameters in Table 1.1 were used as fixed values, and the population parameters (mean vector and covariance matrix) for the latent variables θ_1 and θ_2 were estimated by maximum likelihood. Estimation of the MIRT parameters and subsequent computation of the scale scores was done using the IRTPRO software (Cai et al. 2011).

For the Anxiety scales, the estimated covariance matrix from the fixed (original calibration) parameters and the current data C, with $\theta_1 \equiv \theta_{\text{ped}}$ and $\theta_2 \equiv \theta_{\text{ad}}$, is

$$\hat{\boldsymbol{\Sigma}}_C = \begin{bmatrix} 1.650(0.11) & \\ 1.174(0.07) & 1.047(0.06) \end{bmatrix} . \tag{1.3}$$

The estimated correlation of the two latent variables is $\hat{\rho}_C = \frac{1.174}{\sqrt{1.650 \times 1.047}} = 0.893$. It is convenient to define the ratio of the variance of the adult latent variable to that of the pediatric latent variable, $\hat{k}^2_{\text{ad}} = \frac{1.047}{1.650} = 0.635$.

To compute projected scores on a scale set in a hypothetical calibration population that is the same as the reference population for the pediatric scale, we need an estimate of the covariance matrix of the two latent variables in that population. That estimate has three components: The variance of the pediatric latent variable,

$\hat{\sigma}_{\text{ref(ped),ped}}$, is 1.0 (that set the original scale); the variance of the adult latent variable, $\hat{\sigma}_{\text{ref(ped),ad}}$, is $\hat{k}^2_{\text{ad}}$, obtained from the ratio of the two variances in the current data; and the covariance is $\hat{\rho}_C \hat{k}_{\text{ad}}$, using the correlation from the current data. So the covariance matrix used to project from the pediatric scale to the adult scale is:

$$\hat{\boldsymbol{\Sigma}}_{\text{ref(ped)}} = \begin{bmatrix} 1.0 & \\ \hat{\rho}_C \hat{k}_{\text{ad}} & \hat{k}^2_{\text{ad}} \end{bmatrix} = \begin{bmatrix} 1.000 & \\ 0.711 & 0.635 \end{bmatrix} . \tag{1.4}$$

We also need to compute the predicted value of the adult scale mean in that population. We use linear regression to compute that estimate, based on $\hat{\boldsymbol{\Sigma}}_C$ and the estimated mean vector for the current data, which in this example is

$$\hat{\boldsymbol{\mu}}_C = \begin{bmatrix} 0.631(0.05) \\ 0.868(0.04) \end{bmatrix} . \tag{1.5}$$

The regression estimate of the adult scale value for the pediatric scale mean of 0.0 uses an estimate of the slope

$$\beta_1 = \hat{\rho}_C \frac{\hat{\sigma}_{C,\text{ad}}}{\hat{\sigma}_{C,\text{ped}}} = 0.893 \frac{\sqrt{1.047}}{\sqrt{1.650}} = 0.711 \ , \tag{1.6}$$

and the intercept

Table 1.1 Item parameters for the PROMIS pediatric and adult Anxiety scales, based on their original calibrations

Item	Label	a_1	a_2	c_1	c_2	c_3	c_4
1	Pediatric-Anxiety1-8	1.51	0	1.29	−0.27	−2.80	−4.31
2	Pediatric-Anxiety2-2	1.89	0	0.48	−1.12	−3.24	−4.76
3	Pediatric-Anxiety2-9	1.81	0	1.42	−0.45	−2.87	−4.79
4	Pediatric-Anxiety2-1	1.71	0	0.74	−0.88	−3.00	−4.54
5	Pediatric-Anxiety2-6	1.50	0	0.60	−0.76	−2.78	−3.97
6	Pediatric-Anxiety1-7	1.48	0	1.01	−0.43	−2.84	−4.25
7	Pediatric-Anxiety1-3	1.84	0	0.44	−0.89	−2.83	−4.08
8	Pediatric-Anxiety2-4	1.83	0	−0.46	−1.67	−3.34	−4.69
9	Adult-EDANX01	0	3.60	−1.23	−3.92	−7.06	−9.72
10	Adult-EDANX40	0	3.88	−1.89	−4.91	−8.20	−11.26
11	Adult-EDANX41	0	3.66	−1.33	−3.78	−6.52	−9.59
12	Adult-EDANX53	0	3.66	0.85	−2.18	−5.72	−9.14
13	Adult-EDANX46	0	3.40	0.74	−2.15	−5.59	−9.28
14	Adult-EDANX07	0	3.55	−1.92	−3.71	−6.62	−8.47
15	Adult-EDANX05	0	3.36	0.64	−2.01	−5.28	−8.21
16	Adult-EDANX54	0	3.35	1.71	−1.04	−4.19	−7.69

$$\beta_0 = \hat{\mu}_{C,\mathrm{ad}} - \beta_1 \times \hat{\mu}_{C,\mathrm{ped}} = 0.868 - 0.711 \times 0.631 = 0.419. \tag{1.7}$$

Assuming the same relationships between the pediatric and adult scales observed in the current data *C* would have held if the calibration sample for the pediatric scale had also been the reference sample for the adult scale, the mean for the adult scale would have been

$$\hat{\mu}_{\mathrm{ad}} = 0.419 + 0.711 \times 0.0 = 0.419 \ . \tag{1.8}$$

Assembling all this, calibrated projection of pediatric item responses onto the adult scale uses the item parameters for the pediatric items in Table 1.1, the population mean vector

$$\hat{\boldsymbol{\mu}}_{\mathrm{ref(ped)}} = \begin{bmatrix} 0.0 \\ \hat{\mu}_{\mathrm{ad}} \end{bmatrix} = \begin{bmatrix} 0.0 \\ 0.419 \end{bmatrix} \ , \tag{1.9}$$

and covariance matrix $\hat{\boldsymbol{\Sigma}}_{\mathrm{ref(ped)}}$ from Eq. (1.4).

To project item responses onto the pediatric scale, the computations in this section are reflected appropriately, reversing the roles of the pediatric and adult latent variables.

1.3 A Linear Approximation to Calibrated Projection

In calibrated projection, the projected score on the θ_2 dimension, given the score on the θ_1 dimension, is computed using two-dimensional numerical integration of the conditional posterior distribution, two of which are illustrated in Fig. 1.1. However, it is numerically the case that the predictions so-computed are, within rounding error, a linear function of the predictor scores, specifically

$$\widehat{\mathrm{EAP}}[\theta_2] = \beta_0 + \beta_1 \mathrm{EAP}[\theta_1] \ . \tag{1.10}$$

in which the values of β_0 and β_1 are computed as described in the previous section. So for the Anxiety examples, we can compute

$$\begin{aligned} \widehat{\mathrm{EAP}}[\theta_2] &= \beta_0 + \beta_1 \mathrm{EAP}[\theta_1] \qquad (1.11) \\ &= 0.419 + 0.711 \mathrm{EAP}[\theta_1]. \end{aligned}$$

This linear relationship is exact, due to the linearity of conditional expectations. However, no exact relationship has thus far been found for the values of SD[θ_2]. An approximation that appears empirically useful combines two sources of variance: the error variance of the predicting value, $\mathrm{SD}^2[\theta_1]$, and the residual variance around the regression line, $V_{\mathrm{Res},2}$. The residual variance is

$$V_{\text{Res},2} = (1 - \hat{\rho}_C^2)\hat{\sigma}_{C,2}^2 \tag{1.12}$$

so for the projection from the pediatric scale to the adult scale it is

$$V_{\text{Res},2} = (1 - 0.893^2)1.047 = 0.212 \ . \tag{1.13}$$

Using these values, approximate conditional standard errors of the projected values can be computed for the projection from the pediatric scale to the adult scale as

$$\widehat{\text{SD}}[\theta_2] = \sqrt{\beta_1^2 \text{SD}^2[\theta_1] + V_{\text{Res},2}} = \sqrt{0.711^2 \text{SD}^2[\theta_1] + 0.212} \ . \tag{1.14}$$

1.3.1 Comparing the Results from the Linear Approximation with Calibrated Projection for PROMIS Anxiety

It is not convenient to show the results of projection for scores based on response patterns, because there are so many. However, tabulation of the responses for summed scores covers the entire range and can be useful to provide illustration and checks on the results. The first five columns of Table 1.2 illustrate calibrated projection for the summed scores for the pediatric Anxiety scale, with projection to the adult Anxiety scale. All results are shown using the T-score scale common to all PROMIS measures.

The pediatric EAP[θ_1] and SD[θ_1] values in Table 1.2 are those published as the scoring table for the Anxiety measure. The adult EAP[θ_2] and SD[θ_2] are those computed using two-dimensional quadrature, the item parameters in Table 1.1, and the population mean vector and covariance matrix from Eqs. (1.4) and (1.9). The rightmost four columns of Table 1.2 show the results obtained with the linear approximation described in the preceding section. Columns 6 and 7 show the values of adult $\widehat{\text{EAP}}[\theta_2]$ computed using equation 1.11, and the difference between the calibrated projection EAPs and the linear approximation. Columns 8 and 9 show the values of adult $\widehat{\text{SD}}[\theta_2]$, and the ratio of the approximation to the calibrated projection values. In this case the values from the linear approximation are about 1.2 times larger than those from calibrated projection; but most would still round to the same integral values on the T-score scale.

1.3.2 Summary of Comparisons for Seven PROMIS Scales

In the course of a project to link some of the pediatric PROMIS scales to their adult counterparts, we have computed the results for calibrated projection and the linear approximation described in the preceding section for seven scales. Because all were done twice, once from the pediatric items to the adult scales and a second time from

Table 1.2 The first five columns show the calibrated projection IRT scores (EAPs) and standard errors (SDs) for summed scores on the pediatric measure (only even summed scores are shown to save space)

Pediatric summed score	Calibrated projection				Linear approximation			
	Pediatric		Adult					
	$\mathrm{EAP}[\theta_1]$	$\mathrm{SD}[\theta_1]$	$\mathrm{EAP}[\theta_2]$	$\mathrm{SD}[\theta_2]$	$\widehat{\mathrm{EAP}}[\theta_2]$	$\mathrm{EAP}[\theta_2]-\widehat{\mathrm{EAP}}[\theta_2]$	$\widehat{\mathrm{SD}}[\theta_2]$	$\widehat{\mathrm{SD}}[\theta_2]/\mathrm{SD}[\theta_2]$
0	32.3	5.8	41.7	5.5	41.6	0.0	6.2	1.1
2	39.2	4.7	46.5	4.9	46.5	0.0	5.7	1.2
4	43.3	4.2	49.5	4.7	49.5	0.0	5.5	1.2
6	46.7	3.9	51.8	4.6	51.8	0.0	5.4	1.2
8	49.6	3.8	53.9	4.5	53.9	0.0	5.3	1.2
10	52.3	3.7	55.8	4.5	55.8	0.0	5.3	1.2
12	54.8	3.7	57.6	4.5	57.6	0.0	5.3	1.2
14	57.3	3.7	59.3	4.5	59.4	0.0	5.3	1.2
16	59.7	3.7	61.1	4.5	61.1	0.0	5.3	1.2
18	62.1	3.7	62.8	4.5	62.8	0.0	5.3	1.2
20	64.5	3.7	64.5	4.5	64.5	0.0	5.3	1.2
22	67.0	3.7	66.3	4.5	66.3	0.0	5.3	1.2
24	69.6	3.7	68.1	4.5	68.1	0.0	5.3	1.2
26	72.3	3.7	70.0	4.5	70.0	0.0	5.3	1.2
28	75.2	3.8	72.0	4.5	72.1	−0.1	5.3	1.2
30	78.6	4.1	74.5	4.6	74.5	−0.1	5.4	1.2
32	83.5	4.7	78.0	4.9	78.0	−0.1	5.7	1.2

The last four columns show the results obtained with the linear approximation

the adult item responses to the pediatric scales, there are a total of 14 examples. The latent variables measured by all of the pairs of scales are highly correlated; correlations ranged from 0.86 to 0.95. For all 14 linkings, the linearly approximated EAPs for each summed score were essentially identical to those obtained with numerical integration in calibrated projection, as was illustrated for the Anxiety pediatric to adult projection in Table 1.2.

The degree to which $\widehat{\mathrm{SD}}[\theta_2]$ approximates $\mathrm{SD}[\theta_2]$ as computed by numerical integration in calibrated projection remains an empirical question. While it is not feasible to check that for all response pattern scores, it is easy to evaluate the approximation for the posterior standard deviations associated with each summed score on the scale that is used for projection. An example is shown in Table 1.2, in which the ratio of $\widehat{\mathrm{SD}}[\theta_2]$ to $\mathrm{SD}[\theta_2]$ varies only between 1.1 and 1.2. Table 1.3 shows the minimum and maximum values of that ratio for all 14 of the PROMIS pediatric–adult projections. Across 13 of the 14 cases, the ratio is between 1.0 and 1.3. For many applications, reported standard errors that are zero to 30 % larger than the "exact" values would probably present no problems. The exception is the projection of the Upper Extremity scale from the pediatric to the adult measure, for which $\widehat{\mathrm{SD}}[\theta_2]$ is 1.3–1.6 times larger than $\mathrm{SD}[\theta_2]$.

Table 1.3 The minimum and maximum values of the ratio $\widehat{\mathrm{SD}}[\theta_2]/\mathrm{SD}[\theta_2]$ for all 14 of the PROMIS pediatric-adult projections

	Pediatric to adult		Adult to pediatric	
Pediatric domain	Minimum	Maximum	Minimum	Maximum
Anxiety	1.1	1.2	1.0	1.0
Depressive symptoms	1.1	1.2	1.1	1.1
Anger	1.2	1.3	1.1	1.2
Fatigue	1.3	1.4	1.1	1.1
Pain interference	1.2	1.3	1.0	1.0
Physical functioning—mobility	1.1	1.2	1.0	1.1
Physical functioning—upper extremity	1.3	1.6	1.0	1.3

Table 1.4 Proportions of values of $\mathrm{EAP}[\theta_2]$ for each scale combination that are within ±1 SD and ±2 SD of the values obtained using the linear approximation to calibrated projection

	Pediatric to adult		Adult to pediatric	
Pediatric domain	±1 SD	±2 SD	±1 SD	±2 SD
Anxiety	0.69	0.93	0.72	0.92
Depressive symptoms	0.70	0.93	0.71	0.93
Anger	0.72	0.95	0.74	0.94
Fatigue	0.71	0.94	0.75	0.94
Pain interference	0.75	0.92	0.77	0.95
Physical functioning—mobility	0.71	0.93	0.69	0.91
Physical functioning—upper extremity	0.54	0.95	0.53	0.95

The point of reporting values of $\widehat{\mathrm{SD}}[\theta_2]$ (or $\mathrm{SD}[\theta_2]$) as standard errors for the IRT scale scores is to provide a confidence interval the covers the true value $100(1-\alpha)$ % of the time. With the data collected for linking the pediatric and adult scales, we have the values of $\mathrm{EAP}[\theta_2]$ for each projection, so we can compute the proportion of those values that are included in any specified confidence range. Table 1.4 shows those proportions for confidence intervals computed as $\widehat{\mathrm{EAP}}[\theta_2] \pm \widehat{\mathrm{SD}}[\theta_2]$ and $\widehat{\mathrm{EAP}}[\theta_2] \pm 2\widehat{\mathrm{SD}}[\theta_2]$, which should be about 0.68 and 0.95, respectively, if the standard errors are nearly correct and the errors are approximately normal. Across 13 of the 14 cases, the ±1 SD proportions are between 0.69 and 0.77, while the ±2 SD proportions are between 0.91 and 0.95. While the ±1 SD proportions tend to be a little too large, the ±2 SD proportions are slightly too small. So no improvement could be made on one (e.g., making the SD smaller to reduce the ±1 SD proportions) without making the other worse (the example would make the ±2 SD proportions too small).

The exceptional values in Table 1.4 are the ±1 SD proportions for the Upper Extremity scales, which are 0.53–0.54 instead of 0.68. This anomaly is due to a distributional peculiarity for those scales: In these data, 21 % of the respondents have a perfect (maximum) score on both scales. That produces a single point mass in the distributions with 21 % of the data. The fact that these large blocks of 21 % have residuals between 1 and 2 SDs from zero reduces the observed proportion within ±1 SD from the nominal 0.68 to 0.53–54. Setting that anomaly aside, the coverage proportions suggest that the approximation works well across the linkings that have used it thus far. For future use, it is easy to check its accuracy, by constructing a table like Table 1.2, and comparing the calibrated projection computations for summed scores with the approximation.

1.4 Projection from Two Scales to One, Illustrated with PROMIS Physical Functioning

1.4.1 Calibrated Projection from Two Scales

In this section we extend calibrated projection to use a MIRT model for item responses to three scales. Two measures form the basis of the projection, with θ_1 representing the underlying construct measured by the first scale, with estimated slopes a_1 for each of the first scale's items and fixed values of 0.0 for the other items, and θ_2 representing the underlying construct measured by the second scale, with estimated slopes a_2 for each of the second scale's items and fixed values of 0.0 for the other items. θ_3 represents the underlying construct measured by the third scale, the target scale of the projection, with estimated slopes a_3 for each of the third scale's items and fixed values of 0.0 for the other items. The correlations among all three θs are estimated.

The context for this extension of calibrated projection, and the linear approximation, involves the linking between the PROMIS pediatric Physical Function (PF) scales (Mobility and Upper Extremity/Dexterity; DeWitt et al. 2011) and the adult PF scale (Fries et al. 2014). The results for the Physical Function scales in Tables 1.3 and 1.4 were produced by linking the two pediatric scales separately to the omnibus adult scale; in this section, we will link the two pediatric scales jointly with the adult scale. To do so, we will use the published calibration item parameters for the three unidimensional scales in Table 1.5, where they are expressed as components of a three-dimensional MIRT model, in which θ_1 is pediatric PF-Mobility, θ_2 is pediatric PF-Upper Extremity/Dexterity, and θ_3 is adult PF.

For the PF scales, the estimated covariance matrix from fixed parameters and the current data C, with $\theta_1 \equiv \theta_{\text{ped-Mobility}}$, $\theta_2 \equiv \theta_{\text{ped-UpperExtremity}}$, and $\theta_3 \equiv \theta_{\text{ad-PF}}$, is

$$\hat{\boldsymbol{\Sigma}}_C = \begin{bmatrix} \hat{\boldsymbol{\Sigma}}_{\theta_1,\theta_2} & \hat{\boldsymbol{\Sigma}}_{\theta_{1-2},\theta_3} \\ \hat{\boldsymbol{\Sigma}}'_{\theta_{1-2},\theta_3} & \hat{\sigma}^2_{\theta_3} \end{bmatrix} = \begin{bmatrix} 1.548(0.02) & & \\ 2.189(0.03) & 3.393(0.08) & \\ 1.286(0.07) & 1.841(0.09) & 1.200(0.10) \end{bmatrix} . \tag{1.15}$$

The estimated correlation matrix among the three latent variables is

Table 1.5 Item parameters for the PROMIS pediatric and adult Physical Function (PF) scales, based on their original calibrations

Item	Label	a_1	a_2	a_3	c_1	c_2	c_3	c_4
1	Pediatric-PF-Mobility1-3	3.11	0.00	0.00	5.95	5.42	3.55	1.40
2	Pediatric-PF-Mobility3-9	2.62	0.00	0.00	8.32	6.72	5.29	2.65
3	Pediatric-PF-Mobility4-4	1.96	0.00	0.00	5.69	4.72	3.20	0.97
4	Pediatric-PF-Mobility4-8	3.27	0.00	0.00	10.18	8.67	6.36	4.31
5	Pediatric-PF-Mobility4-3	3.00	0.00	0.00	8.28	7.65	5.78	4.29
6	Pediatric-PF-Mobility2-7	1.82	0.00	0.00	5.78	4.53	3.46	1.73
7	Pediatric-PF-Mobility2-4	1.97	0.00	0.00	5.51	4.73	3.87	2.53
8	Pediatric-PF-Mobility1-1	2.36	0.00	0.00	5.55	4.66	3.15	1.17
9	Pediatric-PF-UpperExtremity2-3	0.00	2.33	0.00	7.63	5.85	3.78	–
10	Pediatric-PF-UpperExtremity4-1	0.00	1.67	0.00	6.45	4.97	3.79	1.26
11	Pediatric-PF-UpperExtremity3-11	0.00	2.53	0.00	7.32	6.97	6.02	3.82
12	Pediatric-PF-UpperExtremity4-10	0.00	1.89	0.00	6.90	5.58	4.54	2.31
13	Pediatric-PF-UpperExtremity3-4	0.00	2.67	0.00	8.99	6.60	4.74	–
14	Pediatric-PF-UpperExtremity3-9	0.00	2.25	0.00	6.59	5.62	4.30	1.56
15	Pediatric-PF-UpperExtremity2-2	0.00	2.54	0.00	10.00	8.20	7.36	4.68
16	Pediatric-PF-UpperExtremity3-7	0.00	2.46	0.00	7.11	6.77	5.37	3.67
17	Adult-PFA1	0.00	0.00	3.31	3.71	1.66	−0.43	−1.99
18	Adult-PFC36	0.00	0.00	4.46	6.38	4.50	2.63	1.03
19	Adult-PFC37	0.00	0.00	4.46	10.30	7.27	4.68	2.54
20	Adult-PFA5	0.00	0.00	4.14	9.81	6.71	4.31	2.19
21	Adult-PFA3	0.00	0.00	2.95	6.64	3.72	1.65	−0.09
22	Adult-PFA11	0.00	0.00	4.83	9.56	7.39	5.36	2.17
23	Adult-PFA16	0.00	0.00	3.37	10.58	8.63	6.44	4.18
24	Adult-PFB26	0.00	0.00	3.32	10.52	9.56	7.77	5.84
25	Adult-PFA55	0.00	0.00	3.58	11.99	9.49	7.41	5.30
26	Adult-PFC45	0.00	0.00	3.11	9.67	8.65	6.87	4.54

Note: For two of the Pediatric Upper Extremity items, two response categories were collapsed in calibration so there are only three intercepts for those items

$$\hat{\mathbf{R}}_C = \begin{bmatrix} 1.000 & & \\ 0.955 & 1.000 & \\ 0.944 & 0.912 & 1.000 \end{bmatrix}, \tag{1.16}$$

and the estimated mean vector for the current data is

$$\hat{\boldsymbol{\mu}}_C = \begin{bmatrix} \hat{\boldsymbol{\mu}}_{\theta_1,\theta_2} \\ \hat{\mu}_{\theta_3} \end{bmatrix} = \begin{bmatrix} -0.634(0.06) \\ -0.269(0.10) \\ -0.332(0.05) \end{bmatrix}. \tag{1.17}$$

To compute estimates of the mean and covariance matrix among the latent variables for a hypothetical joint reference distribution for the pediatric and adult scales, and for use in the linear approximation to calibrated projection, we need the regression coefficients for θ_3 on θ_1 and θ_2, which are

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \widehat{\beta_1} \\ \widehat{\beta_2} \end{bmatrix} = \hat{\boldsymbol{\Sigma}}^{-1}_{\theta_1,\theta_2} \hat{\boldsymbol{\Sigma}}_{\theta_{1-2},\theta_3} = \begin{bmatrix} 0.724 \\ 0.076 \end{bmatrix} \; ; \tag{1.18}$$

the intercept is

$$\hat{\beta}_0 = \hat{\mu}_{\theta_3} - \hat{\boldsymbol{\beta}}' \hat{\boldsymbol{\mu}}_{\theta_1,\theta_2} = 0.147 \; . \tag{1.19}$$

To obtain the mean vector for the hypothetical joint reference distribution for the pediatric and adult scales, we proceed as we did in Sect. 1.2, and compute the predicted value of the adult mean from the regression equation,

$$\hat{\mu}_{\text{ad}} = \beta_0 + \beta_1 \times 0.0 + \beta_2 \times 0.0 = 0.147 + 0.724 \times 0.0 + 0.076 \times 0.0 = 0.147 \; , \tag{1.20}$$

so the mean vector used to project from the pediatric scale to the adult scale is:

$$\hat{\boldsymbol{\mu}}_{\text{ref(ped)}} = \begin{bmatrix} 0.000 \\ 0.000 \\ 0.147 \end{bmatrix} \; . \tag{1.21}$$

Assembly of an estimate of the covariance matrix for the hypothetical pediatric-adult reference population is more challenging than it was in the two-dimensional case described in Sect. 1.2. However, we proceed with a similar series of steps: (a) We set the variances of the two pediatric measures, θ_1 and θ_2, to the reference value of 1.0. (b) We use the estimate of the correlation between θ_1 and θ_2 obtained from the current data. (c) We compute a proportionally adjusted estimate of the variance for the adult θ_3. (d) Finally, we combine the estimates of the correlations of θ_1 and θ_2 with θ_3 from the current data with the estimate of the variance of θ_3 to obtain the covariances.

The new challenge appears in step (c): In Sect. 1.2 we used the ratio of the adult variance to the pediatric variance as the adjustment factor $\hat{k}^2$; however, here there are two pediatric variances. In principle, it might be possible to use a ratio constructed from any combination of the two pediatric variances. In this illustration we use the weighted combination that is the regression prediction of adult θ_3 from pediatric θ_1 and θ_2. We compute the variance of the predictions of adult θ_3 in the current data as

$$\hat{\upsilon}^2_C = \beta_1^2 \hat{\sigma}^2_{C;\theta_1} + \beta_2^2 \hat{\sigma}^2_{C;\theta_2} + 2\beta_1\beta_2\hat{\sigma}_{C;\theta_1,\theta_2} = 1.070 \; , \tag{1.22}$$

in which the variances and covariance are obtained from the upper left-hand block of the matrix in Eq. (1.15). We compute the (hypothetical) variance of predictions in the reference pediatric sample as

$$\hat{\upsilon}^2_Z = \beta_1^2 + \beta_2^2 + 2\beta_1\beta_2\hat{\rho}_{C;\theta_1,\theta_2} = 0.634 \; . \tag{1.23}$$

Then the predicted variance of adult θ_3 is computed as

$$\hat{k}_2^2\hat{\sigma}^2_{\theta_3} = \frac{\hat{\upsilon}_Z^2}{\hat{\upsilon}_C^2}\hat{\sigma}^2_{\theta_3} = \frac{0.634}{1.070}1.200 = 0.711 \quad . \tag{1.24}$$

Combining that estimate of the adult variance with the correlations in Eq. (1.16) completes the covariance matrix used to project from the pediatric scale to the adult scale:

$$\hat{\boldsymbol{\Sigma}}_{\text{ref(ped)}} = \begin{bmatrix} 1 & & \\ \hat{\rho}_{\theta_1,\theta_2} & 1 & \\ \hat{\rho}_{\theta_1,\theta_3}\hat{k}_{2,}\hat{\sigma}_{\theta_3} & \hat{\rho}_{\theta_2,\theta_3}\hat{k}_2\hat{\sigma}_{\theta_3} & \hat{k}_2^2\hat{\sigma}^2_{\theta_3} \end{bmatrix} = \begin{bmatrix} 1.000 & & \\ 0.955 & 1.000 & \\ 0.796 & 0.769 & 0.711 \end{bmatrix} \quad . \tag{1.25}$$

Calibrated projection, using the item parameters in Table 1.5 and the population mean vector and covariance matrix in Eqs. (1.21) and (1.25) yields the results in the first seven columns of Table 1.6 for the summed scores of pediatric PF Mobility and Upper Extremity/Dexterity combined.

1.4.2 Linear Approximation, from Two Scales to One

To compute the linear prediction of the θ_3 score, we use the regression equation with coefficients from Eqs. (1.18) and (1.19),

$$\begin{aligned} \widehat{\text{EAP}}[\theta_3] &= \beta_0 + \beta_1\text{EAP}[\theta_1] + \beta_2\text{EAP}[\theta_2] \\ &= 0.147 + 0.724\text{EAP}[\theta_1] + 0.076\text{EAP}[\theta_2] \end{aligned} \tag{1.26}$$

and the EAP estimates for θ_1 and θ_2 as the predictor values.

Using reasoning analogous to that expressed in Eqs. (1.12)–(1.14), we compute the residual variance from the regression equation as

$$V_{\text{Res}} = \hat{\sigma}^2_{\theta_3} - \hat{\upsilon}_C^2 = 1.200 - 1.070 = 0.130 \quad , \tag{1.27}$$

and the estimated posterior standard deviations as

$$\begin{aligned} \widehat{\text{SD}}[\theta_3] &= \sqrt{\beta_1^2\text{SD}^2[\theta_1]+\beta_2^2\text{SD}^2[\theta_2]+2(\beta_1\beta_2\text{Cov}[\theta_1,\theta_2])+V_{\text{Res}}} \\ &= \sqrt{0.724^2\text{SD}^2[\theta_1]+0.076^2\text{SD}^2[\theta_2]+2(0.724\times 0.076\text{Cov}[\theta_1,\theta_2])+0.130} \end{aligned} \tag{1.28}$$

in which $\text{Cov}[\theta_1, \theta_2]$ is the error covariance associated with $\text{EAP}[\theta_1]$ and $\text{EAP}[\theta_2]$, with the results shown in the rightmost four columns of Table 1.6.

Table 1.6 The first seven columns show the calibrated projection IRT scores (EAPs) and standard errors (SDs) for summed scores on the pediatric measures combined (only even summed scores are shown to save space)

	Calibrated projection						Linear approximation			
	Pediatric				Adult					
SS	EAP[θ_1]	SD[θ_1]	EAP[θ_2]	SD[θ_2]	EAP[θ_3]	SD[θ_3]	$\widehat{\text{EAP}}[\theta_3]$	d	$\widehat{\text{SD}}[\theta_3]$	r
0	8.7	3.8	7.9	3.5	18.3	4.0	18.4	−0.1	4.7	1.2
2	12.1	3.4	11.6	3.1	21.1	3.8	21.1	−0.1	4.5	1.2
4	14.2	3.2	13.8	2.9	22.7	3.7	22.8	−0.1	4.4	1.2
6	15.8	3.0	15.6	2.8	24.1	3.6	24.1	0.0	4.3	1.2
8	17.3	2.9	17.1	2.7	25.2	3.5	25.3	0.0	4.2	1.2
10	18.5	2.8	18.4	2.6	26.3	3.5	26.3	0.0	4.2	1.2
12	19.7	2.7	19.6	2.5	27.2	3.4	27.2	0.0	4.2	1.2
14	20.7	2.7	20.7	2.5	28.0	3.4	28.1	0.0	4.1	1.2
16	21.8	2.6	21.7	2.5	28.9	3.4	28.9	0.0	4.1	1.2
18	22.7	2.6	22.7	2.5	29.6	3.4	29.7	0.0	4.1	1.2
20	23.7	2.6	23.7	2.5	30.4	3.4	30.4	0.0	4.1	1.2
22	24.6	2.5	24.6	2.5	31.1	3.3	31.1	0.0	4.1	1.2
24	25.5	2.5	25.6	2.5	31.8	3.3	31.9	0.0	4.1	1.2
26	26.4	2.5	26.5	2.5	32.6	3.3	32.6	0.0	4.1	1.2
28	27.3	2.5	27.4	2.5	33.3	3.3	33.3	0.0	4.1	1.2
30	28.2	2.4	28.3	2.6	34.0	3.3	34.0	0.0	4.0	1.2
32	29.1	2.4	29.2	2.6	34.7	3.3	34.7	0.0	4.0	1.2
34	30.0	2.4	30.2	2.6	35.5	3.3	35.5	0.0	4.0	1.2
36	30.9	2.4	31.1	2.6	36.2	3.3	36.2	0.0	4.0	1.2
38	31.9	2.4	32.1	2.7	37.0	3.3	37.0	0.0	4.0	1.2
40	32.9	2.4	33.1	2.7	37.8	3.3	37.8	0.0	4.0	1.2
42	33.9	2.4	34.1	2.8	38.6	3.3	38.6	0.0	4.1	1.2
44	35.0	2.5	35.2	2.8	39.5	3.3	39.5	0.0	4.1	1.2
46	36.1	2.5	36.3	2.9	40.4	3.4	40.4	0.0	4.1	1.2
48	37.4	2.6	37.6	3.0	41.4	3.4	41.4	0.0	4.1	1.2
50	38.7	2.6	38.9	3.1	42.5	3.4	42.5	0.0	4.1	1.2
52	40.2	2.8	40.4	3.2	43.7	3.5	43.7	0.0	4.2	1.2
54	42.0	2.9	42.1	3.3	45.1	3.6	45.1	0.0	4.3	1.2
56	44.1	3.2	44.2	3.6	46.8	3.7	46.8	0.0	4.4	1.2
58	46.8	3.7	46.9	4.1	49.0	4.0	49.0	0.0	4.6	1.2
60	50.6	4.2	50.5	4.5	51.9	4.3	51.9	0.0	4.9	1.1
62	59.8	6.5	59.7	6.6	59.3	5.9	59.3	0.0	6.3	1.1

The final four columns show the results obtained with the linear approximation. SS is the summed score on the pediatric scales, $d = \text{EAP}[\theta_3] - \widehat{\text{EAP}}[\theta_3]$, and $r = \widehat{\text{SD}}[\theta_3]/\text{SD}[\theta_3]$

As was the case with the one-to-one calibrated projections and their approximations, the values of the linear approximation $\widehat{\mathrm{EAP}}[\theta_3]$ in Table 1.6 are essentially within rounding error of the numerically integrated values $\mathrm{EAP}[\theta_3]$. The ratios of the approximate standard errors $\widehat{\mathrm{SD}}[\theta_3]$ to $\mathrm{SD}[\theta_3]$ are between 1.1 and 1.2, as they were in the one-to-one projections. When the linear approximation values $\widehat{\mathrm{EAP}}[\theta_3]$ and $\widehat{\mathrm{SD}}[\theta_3]$ are combined to produce confidence-interval estimates for the values of response-pattern $\mathrm{EAP}[\theta_3]$ for each respondent in the current data, the proportions covered by the ± 1 SD and ± 2 SD intervals are 0.70 and 0.92, respectively, the former slightly exceeding the target value of 0.68 while the latter is slightly less than the target 0.95, exactly as observed in the one-to-one projections.

In the case of two-to-one projection, the linear approximation does not yield the level of computational simplicity that it did for one-to-one projection, because two-dimensional MIRT scoring is required for θ_1 and θ_2, to obtain the error covariance term in equation 1.28. Given that one is required to compute two-dimensional MIRT scores for the predictor scores, it is probably more straightforward to simply use calibrated projection to compute the three-dimensional MIRT scores that include the estimate for θ_3 as well. Nevertheless, the linear approximation remains a potentially useful pedagogical tool.

1.5 Conclusion

While calibrated projection serves effectively to remove the restriction that IRT calibration could hitherto be used only to link scales that measure the same construct, it is also admittedly mysterious to compute scores on one scale using only item responses from another. The linear approximation presented here is easier to implement, because the projected scores are computed as linear combinations of scores on the basis scales. This also makes apparent the use of regression or prediction in the procedure. The standard errors are computed as the square root of a weighted combination of the error variances of the predicting scores, plus a component due to the imprecision of the regression, all of which is very easy to understand.

While the accuracy of the approximation of the standard error estimates described here remains an empirical question, it is easy to check for summed scores for any particular projection by comparing them to values obtained by numerical integration in calibrated projection. Taken as a whole, the combination of calibrated projection and the linear approximation proposed here extends the scope of linking procedures based on IRT.

References

Cai, L., Thissen, D., & du Toit, S. (2011). *IRTPRO Version 2: Flexible, multidimensional, multiple categorical IRT modeling [Computer software manual].* Chicago, IL: Scientific Software International.

DeWitt, E. M., Stucky, B. D., Thissen, D., Irwin, D. E., Langer, M., Varni, J. W., et al. (2011). Construction of the eight item PROMIS Pediatric Physical Function Scales: Built using item response theory. *Journal of Clinical Epidemiology, 64*, 794–804.

Fries, J. F., Witter, J., Rose, M., Cella, D., Khanna, D., & Morgan DeWitt, E. (2014). Item Response Theory (IRT), Computerized Adaptive Testing (CAT), and PROMIS: Assessment of physical function (PF). *Journal of Rheumatology, 41*, 153–158.

Holland, P. W. (2007). Framework and history for score linking. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 5–30). New York, NY: Springer.

Irwin, D., Stucky, B. D., Langer, M. M., Thissen, D., DeWitt, E. M., Lai, J. S., et al. (2010). An item response analysis of the Pediatric PROMIS Anxiety and Depressive Symptoms Scales. *Quality of Life Research, 19*, 595–607.

Pilkonis, P. A., Choi, S. W., Reise, S. P., Stover, A. M., Riley, W. T., & Cella, D. (2011). Item banks for measuring emotional distress from the Patient-Reported Outcomes Measurement Information System (PROMIS™): Depression, anxiety, and anger. *Assessment, 18*, 263–283.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph* (No. 18).

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

Thissen, D., Varni, J. W., Stucky, B. D., Liu, Y., Irwin, D. E., & DeWalt, D. A. (2011). Using the PedsQL™ 3.0 Asthma Module to obtain scores comparable with those of the PROMIS Pediatric Asthma Impact Scale (PAIS). *Quality of Life Research, 20*, 1497–1505.

Chapter 2
The Reliability of Diagnosing Broad and Narrow Skills in Middle School Mathematics with the Multicomponent Latent Trait Model

Susan Embretson, Kristin Morrison, and Hea Won Jun

Abstract The multicomponent latent trait model for diagnosis (MLTM-D; Embretson and Yang, Psychometrika 78:14–36, 2013) is a conjunctive item response model that is hierarchically organized to include broad and narrow skills. A two-stage adaptive testing procedure was applied to diagnose skill mastery in middle school mathematics and then analyzed with MLTM-D. Strong support for the reliability of diagnosing both broad and narrow skills was obtained from both stages of testing using decision confidence indices.

Keywords Diagnostic models • Item response theory • Multidimensional models • Decision confidence reliability

Diagnostic assessment has become increasingly prominent in the last few years (Leighton and Gierl 2007; Rupp et al. 2010). Several explanatory item response theory (IRT) models (i.e., Hensen et al. 2009; von Davier 2008) have been developed using latent classes to assess patterns of skill or attribute possession by examinees. Since the number of classes increases exponentially with the number of skills that are assessed, the models are typically applied to tests with less than ten skills.

However, using high-stakes broad achievement or proficiency tests that may include 20 or 30 skills, to diagnose more specific skills or skill clusters has several potential advantages. First, the content aspect of validity, as explicated in the *Standards for Educational and Psychological Testing* (2014), is supported, since the tests typically represent skills deemed important by expert panels. Second, proficiencies in the skills represented on the tests have practical importance.

The research in this report was partially supported by a Goal 5 (Measurement) grant from the *Institute of Educational Science* Grant R305A100234 to Georgia Institute of Technology, Susan Embretson, Principal Investigator.

S. Embretson (✉) • K. Morrison • H.W. Jun
Georgia Institute of Technology, 654 Cherry Street, Atlanta, GA 30332, USA
e-mail: susan.embretson@psych.gatech.edu

L.A. van der Ark et al. (eds.), *Quantitative Psychology Research*, Springer Proceedings in Mathematics & Statistics 140, DOI 10.1007/978-3-319-19977-1_2

Decisions about examinees, as well as their instructional support systems, are based on the overall test scores. Third, remedial instructional materials may be coordinated with these tests for examinees who are not deemed to achieve mastery. For example, the *Blending Assessment with Instruction Project* (BAIP; 2010) provides online tutorials that are coordinated with achievement tests administered for Grade 3 to Grade 8. Fourth, the diagnostic assessment would be efficient if the broad tests have sufficiently reliable information about skills. However, this last advantage is questionable because commonly used subscale scores often do not have sufficient reliability (Sinharay 2010) particularly when the subscales are highly intercorrelated.

The purpose of this study is to examine the reliability of diagnosis from heterogeneous tests for mastery of skill clusters and specific skills (Fig. 2.1). An example of a two-stage adaptive diagnostic system is presented that was applied to mathematics achievement in middle school. The methods employed differ from using subscale scores in several important ways. First, the study employs a diagnostic IRT model. In the current study, a diagnostic model that is appropriate for a heterogeneous test, the multicomponent latent trait model for diagnosis (MLTM-D; Embretson and Yang 2013), is applied. Second, the broad achievement test is not necessarily viewed as sufficient for diagnosis. Instead, the broad test is Stage 1 in a multistage adaptive testing (MST) design for diagnosis. Stage 2 testing can be adapted to those skill clusters that are not sufficiently reliable in Stage 1. An interesting issue is the extent to which diagnosis may be sufficiently reliable from the Stage 1 heterogeneous test. Third, since the goal is to provide diagnosis, not accurate score locations on a continuum, different indices for reliability may be appropriate. Since diagnosis depends on cutlines, decision accuracy and consistency indices may be applied (Lewis and Sheehan 1990).

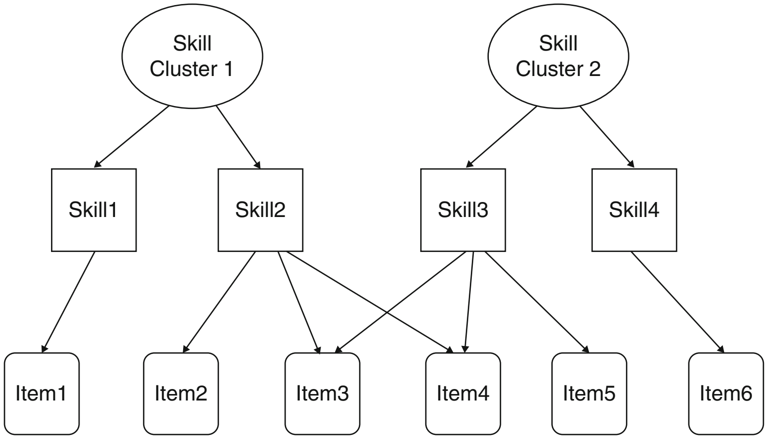

Fig. 2.1 Hierarchical blueprint structure

Prior to presenting results from the two-stage diagnostic testing of mastery of skills in middle school mathematics, an overview of the diagnostic model and procedures, as well as a consideration of appropriate reliability indices, is presented.

2.1 Background

2.1.1 Diagnostic Modeling of Heterogeneous Tests

The diagnostic model. The MLTM-D is a confirmatory model that is appropriate for hierarchically organized test domains with complex items. That is, separately defined areas of competency are represented on the test and more narrowly defined skills are clustered into these broader areas (see Fig. 2.1). Some items may involve skills from only one cluster while other items may involve skills from two or more clusters. To implement MLTM-D, two sets of scores are required: C_{ixk} is the involvement of component k in item i (i.e., the skill cluster), and $Q_{ixm(k)}$ is the score for item i on skill/attribute m with component k. The probability that the response of person j to the total item iT, $\mathrm{X}_{\mathrm{ijT}}$, is correct depends on the probability of solving the relevant skill clusters, as follows:

$$P\left(X_{ijT}=1\right)=\prod\nolimits_{k}P\left(X_{ijk}=1\right)^{cik} \tag{2.1}$$

and

$$P(X_{ijk}=1)=1/(1+\exp(-1.7(\theta_{jk}-\sum_{m}\eta_{km}q_{ikm}+\eta_0))), \tag{2.2}$$

where $\mathrm{X}_{\mathrm{ijk}}$ is the response of examinee j to component k on item i, θ_{jk} is the trait level of examinee j on component k, $\mathrm{q}_{\mathrm{ikm}}$ is the score for stimulus feature m in component k for item i, η_{km} is the weight of feature m on component k, and c_{ik} is the involvement of component k in item i. The within component model for MLTM-D is similar to a linear logistic test model (LLTM; Fischer 1973). It should be noted that $\mathrm{X}_{\mathrm{ijk}}$ is not directly observable, but that the associated parameters can be estimated from response patterns in the data.

Setting mastery boundaries in MLTM-D. For skill clusters, mastery levels can be set by locating skills on the components of MLTM-D. These probabilities are often set for the test as a whole by expert panels, but they also may be applied to skills and skill clusters in MLTM-D. Define $\overline{P}_m$ as the mean predicted probability of solving items on component k for θ_k. Then the cutline τ_k for component k may be found so that $\overline{P}_m \geq y$, where y is a specified probability for mastery.

For specific skills, as for component mastery, a probability for mastery, y, also must be specified. Specific attributes or skills are located on the common scales for component traits and items by their parameter estimates. Assuming that skill m is specified by a binary variable q_{km} for each relevant component, the estimated η_{km} indicates skill position on the theta scale where P_{km} equals .50. However, mastery of skill m must be located at a specified probability, y, within each relevant component. That is, the location of skill m in component k, γ_{mk}, is determined by the probability of solving skill m, P_{km}, such that $P_{km} = y$ if $\theta_k = \gamma_k$. Skill mastery for examinee j on skill m in component k is scored as 1 if $\theta_{jk} \geq \gamma_{m(k)}$, otherwise skill mastery is scored as 0. Number of skills mastered for component k is the sum of the mastered skills. It should be noted that interpretability of skill mastery depends on the strength of prediction of item difficulty by the skills involved.

2.1.2 Assessing Reliability and Decision Accuracy

Empirical reliability. Reliability of component estimates in MLTM-D may be obtained by traditional methods, which depend on the how the traits are estimated (see du Toit 2003). For *expected a posteriori* estimates (EAP), assuming the Rasch model specified within components in MLTM-D, empirical reliability for component k is given as:

$$\rho_t = \sigma^2_{\theta k} / \left(\sigma^2_{\theta k} + \overline{\sigma^2_{\varepsilon k}} \right), \tag{2.3}$$

where $\sigma^2_{\theta k}$ and $\overline{\sigma^2_{\varepsilon k}}$ are the variance of θ_k and the mean error variance, respectively, for component k.

Decision accuracy. If MLTM-D component estimates are used for mastery decisions, cutlines are applied as described above, decision accuracy estimates may be more appropriate for describing score properties. Decision accuracy has often been defined in terms of IRT estimates (Lewis and Sheehan 1990; Rudner 2005; Wainer et al. 2005). While these researchers were primarily interested in providing indices for decision accuracy for the test as a whole (not components or skill clusters), the underlying basis of the indices is interesting to consider. Rudner (2005), for example, placed the mastery cutline for the test, τ_w, on the estimated plausible distribution of theta, θ^*_j, for each person, assuming $\theta^*_j \sim N\left(\theta_j, \sigma^2_\varepsilon\right)$. For $\theta_j \geq \tau_w$, the proportion of $\theta^*_j \geq \tau_w$ would indicate accuracy. Conversely, for $\theta_j < \tau_w$, the proportion of $\theta^*_j < \tau_w$ would indicate accuracy. Thus, decision accuracy depends on both distance from the cutline and the standard error of measurement.

Given this formulation of procedures, decision confidence, ζ_j, also can be expressed for each person as follows:

$$\zeta_j = max\left(P^M_j,\ P^{NM}_j\right), \tag{2.4}$$

where P^M_j is probability of mastery (theta equal to or above cutline) and P^{NM}_j is the probability of non-mastery or $1 - P^M_j$. In turn, P^M_j is obtained as follows: