RUNNING HEAD: Multidimensional Item Response Theory

Scoring Subscales using Multidimensional Item Response Theory Models

Christine E. DeMars

James Madison University

Author Note

Christine E. DeMars, Center for Assessment and Research Studies, James Madison University.

Correspondence concerning this manuscript should be addressed to Christine DeMars, Center for Assessment and Research Studies, MSC 6806, James Madison University, Harrisonburg VA 22807.

Abstract

Several methods for estimating item response theory scores for multiple subtests were compared. These methods included two multidimensional item response theory models: a bi-factor model where each subtest was a composite score based on the primary trait measured by the set of tests and a secondary trait measured by the individual subtest, and a model where the traits measured by the subtests were separate but correlated. Composite scores based on unidimensional item response theory, with each subtest borrowing information from the other subtest, as well as independent unidimensional scores for each subtest were also considered. Correlations among scores from all methods were high, though somewhat lower for the independent unidimensional scores. Correlations between course grades and test scores, a measure of validity, were similar for all methods, though again slightly lower for the unidimensional scores. To assess bias and RMSE, data were simulated using the parameters estimated for the correlated factors model. The independent unidimensional scores showed the greatest bias and RMSE; the relative performance of the other three methods varied with the subscale.

Scoring Subscales using Multidimensional Item Response Theory Model

Tests are often designed such that each item measures the primary trait and one additional secondary trait. The secondary traits may reflect different content categories in the test blueprint, or different tests within a battery of tests. In this situation, test users may want subscale scores, each of which reflects both the primary trait and the relevant secondary trait. Two multidimensional item response theory (MIRT) models are potentially useful in this context: a model with *n* correlated traits, where *n* is the number of subscales, or a bi-factor model with one primary trait and *n* orthogonal secondary traits. An additional model, which applies unidimensional IRT in the initial scoring of each subscale but then borrows information from correlated subscales in forming the final subscale scores, could also be applied.

In the bi-factor model, all items are specified to load on the primary factor. Additionally, each item may load on one additional factor. The factors are orthogonal (Gibbons & Hedeker, 1992; McLeod, Swygert, & Thissen, 2001). In other words, a secondary factor is the common factor a group of items shares beyond their shared association with the primary factor. *Hierarchical* is a more general term for this class of models; *bi-factor* emphasizes that each item loads on no more than two traits, including the primary trait. With the bi-factor model, scores can be estimated for the primary trait and each secondary trait. On a battery of tests, though, it would seem desirable for each subtest score to be a measure of the overall construct covered in the subtest, not just the part of the construct not covered by the primary factor. In other words, the score should be a combination of the primary trait and the secondary trait, not just the secondary trait. To quantify the relative weights of the factors contributing to an item response, Reckase (1985; 1997; Reckase & McKinley,1991) defined the direction of greatest slope for item *i* as

$$\alpha_{ik} = \arccos \frac{a_{ik}}{\sum_{k=1}^{m} a_{ik}^2}, \quad (1)$$

where α_{ik} is the angle with axis *k* and a_{ik} is the discrimination parameter for trait *k*. In the bi-factor model, it is simplest to view the angle for each item relative to the primary factor. If an item measured only the primary trait, α would be 0; if an item measured the primary trait and secondary trait equally, α would be 45º. Though each item may measure a slightly different composite of the primary and secondary traits, an average composite could be formed for each subtest, based on the average angle with the primary axis for the items in the subtest.

Multidimensional IRT Approach

TESTFACT (Bock, Gibbons, Schilling, Muraki, Wilson, & Wood, 2003) and NOHARM (Fraser, 1988) both estimate the item parameters for multidimensional normal ogive models for dichotomous items. The model estimated is:

$$P_i(\boldsymbol{\theta}) = c_i + (1 - c_i)\Phi\left(\mathbf{a}_i'\boldsymbol{\theta} + d_i\right), \quad (2)$$

where $P_i(\boldsymbol{\theta})$ is the probability of correct response on item *i* given the $\boldsymbol{\theta}$ vector of abilities and the item parameters, Φ indicates the cumulative standard normal distribution, c_i is the lower asymptote, $\mathbf{a}_i$ is a vector of discrimination parameters, and d_i is the item difficulty. In contrast to the common unidimensional models, d_i is added, not subtracted, so easier items have higher values for d. TESTFACT uses full information maximum likelihood estimation for estimating the parameters (Bock, Gibbons, & Muraki, 1988; Gibbons & Hedeker, 1992; Muraki & Engelhard, 1985), while NOHARM uses bivariate information (proportion correct for each item and joint proportion correct for each pair of items) and estimates a polynomial approximation to the normal ogive model (McDonald, 1997; 1999).

Though both TESTFACT and NOHARM use the normal ogive models, the parameters estimated should be virtually the same as those of the multidimensional logistic IRT model, which may be more familiar to some readers:

$$P_i(\boldsymbol{\theta}) = c_i + (1 - c_i)\frac{e^{1.7(\mathbf{a}_i'\boldsymbol{\theta} + d_i)}}{1 + e^{1.7(\mathbf{a}_i'\boldsymbol{\theta} + d_i)}}, \quad (3)$$

where the parameters are as defined for Equation 2.

TESTFACT and NOHARM parameter estimates have been compared for exploratory models. Zhang and Stone (2004) found that TESTFACT and NOHARM produced very similar parameter estimates when used in an exploratory mode with items measuring two uncorrelated factors. Knol and Berger (1991) recovered the parameters of one-, two-, and three-dimensional exploratory IRT models using TESTFACT, NOHARM, MAXLOG, and traditional linear factor analysis methods. Of the non-linear (IRT) packages, MAXLOG had the largest RMSE between the parameter estimates and true parameters. TESTFACT performed slightly better than NOHARM. Miller (1991) reviewed earlier comparisons of TESTFACT, MIRTE, and MULTIDIM by Ackerman, and using the same data set added a study of NOHARM. The mean and standard deviation of the residuals were similar for NOHARM and TESTFACT, a bit higher for MULTIDIM, and quite large for MIRTE. Examining the recovering of individual item parameter within NOHARM, the standard errors and bias were large for the anchor items, and choosing items with average levels of difficulty and discrimination or items expected to load highly on a given dimension as anchors did not help.

Gosz and Walker (2002) compared the probability of correct response based on the true item parameters and the item parameters estimated in TESTFACT and in NOHARM. True theta values were used with the recovered item parameters in the probability calculations. Items were generated from a two-dimensional extension of the 2PL model; in some conditions one subset of

items loaded only on factor and the remaining items loaded only on factor two, while in other conditions a subset of items loaded on both factors. Exploratory analyses, with two factors, were used to recover the item parameters. The probabilities of correct response based on the NOHARM item parameter estimates were closer to the true probabilities than were those based on the TESTFACT estimates. NOHARM's performance was particularly better than TESTFACT's for items with a higher discrimination value on the second factor.

De Champlain and Gessaroli (1998) tested for unidimensionality using TESTFACT and NOHARM with small samples and short tests. In NOHARM, unidimensionality was assessed by fitting a one-factor model and using a χ^2 fit index based on the standardized residuals of the interitem joint-probability matrix. In TESTFACT, both unidimensional and exploratory two-dimensional models were estimated, and the difference in log-likelihood χ^2 fit was calculated. When the data were unidimensional, TESTFACT had high Type I error rates, while the error rate for NOHARM was close to the nominal value. When the data were two-dimensional, both methods had high power when the traits were uncorrelated or when the test had 40 items, but NOHARM had higher power when the traits were correlated and the test had only 20 items.

None of these studies compared NOHARM and TESTFACT when used for confirmatory analysis of multidimensional models. Also, these studies did not compare different confirmatory models such as the bi-factor model with a model with correlated factors.

Alternative to Multidimensional Models

When several tests are administered to examinees, or when a test includes subscales, the parameters for each subtest or subscale may be estimated separately. In addition to avoiding dependency problems due to items on a subscale sharing an additional factor beyond the primary test trait, this approach allows for estimation of several trait scores. The potential drawback of

this approach is that the subtests may be quite short and unreliable. Wainer et al. (2001) developed a score augmentation approach that uses information from the other subscales in estimating a subscale score. This method can be used with both classical test theory and IRT-based scores. The extent of the influence of the other subscores depends on the intercorrelations of the subscores and their reliabilities. For example, a subtest that had low reliability and was highly correlated with another subscale with high reliability would be affected more. Augmented subscale scores are estimated by:

$$\hat{\boldsymbol{\tau}}_j = \mathbf{x}_\cdot + \mathbf{S}^{true}(\mathbf{S}^{obs})^{-1}(\mathbf{x}_j - \mathbf{x}_\cdot), \quad (4)$$

where $\hat{\boldsymbol{\tau}}_j$ is a vector of augmented subscale scores for examinee j, $\mathbf{x}_j$ is a vector of unaugmented subscale scores for person j and $\mathbf{x}_\cdot$ is a vector of subscale means, and $\mathbf{S}^{true}$ and $\mathbf{S}^{obs}$ are the estimated true and observed variance-covariance matrices. When applied to Bayesian IRT scores (expected a-posterior (EAP) or modal a-posterior (MAP)), the bias of the scores toward the subscale means is first adjusted for, based on the subscale marginal reliability. Each diagonal element of $\mathbf{S}^{obs}$ and $\mathbf{S}^{true}$ is divided by the squared marginal reliability of the corresponding subscale, and each off-diagonal element is divided by the product of the reliabilities of the corresponding subscale. Each subscale score in $\mathbf{x}_j$ is also divided by the marginal reliability. The vector $\mathbf{x}_\cdot$ will be a vector of 0's if the IRT metric has been scaled so that the estimated population mean of each subscale is zero.

Gessaroli (2004) applied Wainer et al's (2001) method to a test with four correlated subscores, using the number-correct scores. He also estimated scores with the bi-factor model; using the thetas and item parameter estimates from the bi-factor model he calculated the expected score on the number-correct metric. Results from the two methods were essentially

identical, and these methods yielded scores with greater reliability/smaller standard errors than raw scores.

Purpose of the study

In this study, estimated scores from the bi-factor model, a multidimensional model with correlated factors, and the augmented score approach were examined. The bi-factor model was estimated in TESTFACT and NOHARM, and the subscale scores were estimated as a weighted combination of the primary factor and the associated subscale factor. An additional model was estimated in NOHARM: the items on each subscale formed a factor and the factors were free to correlate. Separate unidimensional models for each subscale were also estimated, and Wainer et al.'s (2001) data augmentation method was used to estimate scores for each subscale. In addition, the unaugmented unidimensional scores themselves were examined. Note that all of these methods used IRT scores (thetas); unlike Gessaroli's (2004) study the IRT scores were not used to estimate scores in the number-correct metric. These results would be most applicable when standardized test scores are a direct transformation of the IRT thetas.

Method

Participants

Second and third-year students at James Madison University participated in this study. The university requires students with 45-70 credit hours to participate in assessment activities each spring. The students are randomly assigned to assessment instruments covering different areas of general education or student development. Over the course of three years, 2552 students completed the two tests selected for this study.

Instruments and Procedures

Two multiple-choice assessment tests were used for this study, a scale covering American history and political science and a scale covering global issues. Though the tests administered to the students were longer, 20 of the American items and 15 of the global items were used in this study to create a situation where the scales were short enough that reliability would be increased by using a multidimensional model or augmented scores. The tests were presented at the same testing session along with varying other instruments. Tests were not strictly timed but when most students had finished a test the others were given an additional five minutes to finish. Test scores did not appear on student transcripts and were used only in the aggregate for program assessment.

Estimation

Item parameters for the bi-factor model were estimated twice, once using TESTFACT 4.0[1] and once using NOHARM. NOHARM does not provide ability estimates, so EAP (expected a-posteriori) ability estimates were estimated using TESTFACT 4.0, once using the item parameters from TESTFACT and again using the item parameters from NOHARM. Ability was estimated for both the primary and secondary factors (American history was one secondary factor and global issues was the other). Nine evenly spaced quadrature points from -4 to 4 were used for each dimension. For each of the two subtests, the average angle with the primary trait axis was calculated. Based on this angle, the subscale ability was calculated as a weighted linear composite. The ratio of the weight for θ_2 to the weight for θ_1 used in forming the composite score is equal to the tangent of this angle (Ackerman, 1991). While any weights with this ratio could be used, the weights used here were chosen such that the sum of their squares was equal to one; because the θs were uncorrelated and they were scaled such that their estimated variances

were 1, this would result in a composite score with an estimated variance of 1 (the observed variance of the Bayesian ability estimates was of course less than 1 due to shrinkage).

A model with two correlated factors (based on the two subscales) was also estimated in NOHARM. Because TESTFACT only calculates ability estimates for uncorrelated factors, and NOHARM does not provide ability estimates, a routine for EAP estimation was written using SAS. The quadrature points for each dimension were the same as those used in TESTFACT for orthogonal factors, but the prior densities at each point were based on a bivariate normal distribution with a correlation equal to the correlation between the factors estimated in NOHARM.

Wainer et al.'s (2001) data augmentation method uses scores estimated for each subscale separately. BILOG-MG 3.0 (Zimowski, Muraki, Mislevy, & Bock, 2003) was used to estimate a three-parameter logistic (3-PL) unidimensional model for each subscale. EAP scores were estimated for each subscale within BILOG-MG as well. Augmented subscale scores were then calculated following Equation 4 with the adjustments noted for EAP scores.

Results

Bi-factor model. For the bi-factor model, using NOHARM item parameter estimates the average angle with the primary factor axis was 21 degrees for the American subtest and 30 degrees for the global subtest. The angles based on the TESTFACT item parameter estimates were similar: 27 and 25 degrees. Thus, both subscales were measuring the primary dimension more than the secondary dimensions. The resulting weighted composites, based on these average angles, were $\theta_{American} = 0.93\ \theta_1 + 0.36\ \theta_2$ and $\theta_{Global} = 0.86\ \theta_1 + 0.50\ \theta_3$ for NOHARM, and $\theta_{American} = 0.90\ \theta_1 + 0.43\ \theta_2$ and $\theta_{Global} = 0.91\ \theta_1 + 0.42\ \theta_3$ for TESTFACT.

2-Factor model. When the items were calibrated with a 2-factor model in NOHARM, the estimated correlation between the factors was .81. Thus, the prior distribution used for the EAP estimation of the θs was a bivariate normal distribution with a correlation of .81. For each examinee, the posterior distribution of θ_1 and θ_2 was approximated using quadrature methods as described in the *Method* section; the joint distribution was marginalized over each dimension in turn, and the mean of this distribution taken as the EAP score.

Augmented Subscale Scores. The estimated marginal reliability was .76 for the unidimensional American scores and .68 for the global scores. The variances of the score estimates were .77 and .68 (EAP scores are biased inward, so the variances of the estimated scores are lower than the estimated variances of the scores), with a covariance of .42. Substituting these numbers into Equation 4 and adjusting for the bias in the EAP estimates, the functions for the augmented scores were: $\theta_{American(augmented)} = .64/.76\ \theta_{American} + .20/.68\ \theta_{Global}$ and $\theta_{Global(augmented)} = .29/.76\ \theta_{American} + .52/.68\ \theta_{Global.}$

Comparison of Score Estimates. Tables 1 and 2 show the mean, standard deviation, minimum, and maximum of the score estimates. The means for all methods are essentially zero. The independent unidimensional models method led to scores with a smaller range, and somewhat smaller standard deviation in the case of the global subtest; the more extreme scores were pulled towards the mean more with the unidimensional models, presumably because of lower reliability.

Correlations among the scores are shown in Tables 3 and 4. With the exception of the separate unidimensional models, correlations among the scores based on different methods were all at least 0.99. The unidimensional scores had correlations of at least 0.94 with scores from the other methods.

Comparison of Correlations with Course Grades. The objectives on the test blueprints matched the objectives for general education courses. For the American history and political science curriculum, students selected one of two courses: US history or US political science. For global issues, students chose from five courses designed to address the global issues curricular objectives; these courses were offered in the departments of anthropology, economics, geography, political science, and sociology. Correlations between the test scores and the relevant course grades are shown in Tables 5 and 6. Correlations varied depending on the course; within each course, the unidimensional scores consistently had a *slightly* smaller correlation with the course grades.

Comparison of Bias and RMSE. Because real data were used in this example, it is difficult to know which model produces the most accurate estimates. Simulations were run to assess bias and RMSE of the ability estimates. For these simulations, the item parameters estimated from the real data using the two-factor model were used as the generating, or true, item parameters. A sample of 2500 simulees were drawn from a bivariate normal distribute with a correlation of .81 between the abilities. Item responses were simulated for each of 100 replications, using the logistic parameterization in Equation 3 for convenience. Item parameters were recovered for each replication, and the ability parameters were estimated based on the item parameter estimates, not the generating item parameters. Thus, errors in item parameter estimation and errors in estimating the coefficients for the linear combinations used in the bi-factor and augmented score methods were taken into account. The simulation was conducted for the bi-factor model, the 2-factor model, the augmented approach, and the independent unidimensional models; for the bi-factor model, because the scores were virtually identical using TESTFACT and NOHARM, only TESTFACT was used in the simulation.

Figures 1 and 2 show the bias across the ability range. As would be expected with EAP scores, scores were biased toward the mean. For the American subtest, the augmented scores were the least biased and the unidimensional scores were the most biased. Bias was slightly greater for the bi-factor model than for the 2-factor model. For the Global subtest, both the augmented scores and the unidimensional scores were more biased than the bi-factor and 2-factor models, and bias was again slightly greater for the bi-factor model than for the 2-factor model.

Figures 3 and 4 show the RMSE across the ability range. For both subtests, RMSE tended to be greatest for the unidimensional method. For the American subtest, in the center of the ability distribution both the augmented and unidimensional scores had somewhat greater RMSE than the multidimensional IRT approaches. At abilities below -1.5 and above 1.5, RMSE was lowest for the augmented scores and greatest for the unidimensional scores, with RMSE for the bi-factor and 2-factor models in the middle. For the Global subtest, in the center for the distribution RMSE was slightly lower for the augmented scores, but at more extreme scores below -1.5 or above 1.5, RMSE was lower for the bi-factor and 2-factor models. This is opposite the pattern seen for the American subtest. These differing patterns were consistent with the results for bias; the bias for the American scores was lowest for the augmented scores so the augmented scores might be expected to have lower RMSE in the extremes providing they did not have much greater standard deviations. Similarly, the bias for Global scores was lower for the multidimensional IRT approaches, so at the extremes these scores might be expected to have lower RMSE, again assuming they did not have appreciably larger standard deviations.

Discussion, Limitations, and Conclusions

Scores on the American history and political science test and the global issues test used in this study were nearly the same based on the TESTFACT and NOHARM bi-factor models, a 2-correlated-factor model, and the augmented score approach. Correlations among these methods were .99 or higher and the mean, standard deviation, and range of scores were similar. Scores based on two separate unidimensional models had slightly lower, though still high, correlations with the other models, but a smaller standard deviation and range, presumably because they were biased toward the mean more than the scores from the other approaches. Correlations between the scores and course grades were slightly lower for the independent unidimensional model. For any one course the difference was so small it would not be worth mentioning, except that this correlation was consistently the smallest correlation for each course.

The simulation study showed the unidimensional scores were more biased and had higher RMSE. The relative bias and RMSE for the other approaches differed on the two tests. The bi-factor and 2-factor models showed very similar levels of bias and RMSE; on one test higher than the augmented scores at the extremes, and on the other test lower. Based on these results, there is no clear advantage for any of these three methods over the others, but all produced lower bias and RMSE than the separate unidimensional models.

A limitation of using a real data set, or item parameters based on a real data set for the simulation study, is that it is difficult to know how the results will generalize to other tests. On the other hand, with a real data set the results are at least realistic for one situation. These results might generalize to other situations where the subtests are of moderately short length (15-20 items) and measure relatively similar skills and are administered at the same time. They might be

less generalizable to longer or shorter subtests or to those which measure more disparate skills or are administered further apart in time.

The implications for this study are that with tests of 15-20 items, less biased scores with smaller standard errors can be obtained using a multidimensional IRT model or Wainer et al.'s (2001) augmented score approach. The augmented scores might be the least arduous to calculate; for this study BILOG-MG was used for estimating the unidimensional scores and the IML procedure within SAS was used to estimate the coefficients for combining these scores. The correlated-factor model might be most conceptually appealing, but I am aware of no commercial software which will estimate IRT scores for correlated factors[2], so calculation of scores must be done separately and requires a geometrically increasing number of quadrature points as the number of factors increases.

References

Ackerman, T. A. (1991). The use of unidimensional parameter estimates of multidimensional items in adaptive testing. *Applied Psychological Measurement, 15*, 13-23.

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement, 12*, 261-280.

Bock, R. D, Gibbons, R., Schilling, S. G., Muraki, E., Wilson, D. T., & Wood, R. (2003). *TESTFACT 4.0* [Computer software and manual]. Lincolnwood, IL: Scientific Software International.

De Champlain, A., & Gessaroli, M. E. (1998). Assessing the dimensionality of item response matrices with small sample sizes and short test lengths. *Applied Measurement in Education, 11*, 231-253.

Fraser, C. (1988). *NOHARM*. [Computer software and manual]. Armidale, New South Wales, Australia: author.

Gessaroli, M. E. (2004, April). *Using hierarchical multidimensional item response theory to estimate augmented subscores*. Paper presented at the annual meeting of the National Council for Measurement in Education, San Diego, CA.

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information bi-factor analysis. *Psychometrika, 57*, 423-436.

Gosz, J. K., & Walker, C. M. (2002, April). *An empirical comparison of simple vs. complex multidimensional item response data using TESTFACT and NOHARM*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Knol, D. L., & Berger, M. P. F. (1991). Empirical comparison between factor analysis and multidimensional item response models. *Multivariate behavioral research, 26*, 457-477.

McDonald, R. P. (1997). *Normal-ogive multidimensional model*. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (pp. 257-269). New York: Springer.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum Associates.

McLeod, L. D, Swygert, K. A., & Thissen, D. (2001). Factor analysis for items scored in two categories. In D. Thissen and H. Wainer (Eds.), *Test scoring* (pp. 189-216). Mahwah, NJ: Lawrence Erlbaum Associates.

Miller, T. R. (1991). *Empirical estimation of standard errors of compensatory MIRT model parameters obtained from the NOHARM estimation program*. (Report No. ONR-91-2). Iowa City: ACT. (ERIC Document Reproduction Service No. ED 339738)

Muraki, E., & Engelhard, G., Jr., (1985). Full-information item factor analysis: Applications of EAP scores. *Applied Psychological Measurement, 9*, 417-430.

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement, 9*, 401-412.

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.

Reckase, M. D., & McKinley, R. L. (1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement, 15*, 361-373.

Wainer, H., Vevea, J. L., Camacho, F., Reeve, B. B., III, Rosa, K., Nelson, L., et al. (2001). In D. Thissen and H. Wainer (Eds.), *Test scoring* (pp. 189-216). Mahwah, NJ: Lawrence Erlbaum Associates.

Wu, M. L, Adams, R. J., & Wilson, M. R. (1998). *CONQUEST* [Computer software and manual]. Melbourne, Victoria, Australia: The Australian Council for Educational Research.

Zhang, B., & Stone, C. (2004, April). *Direct and indirect estimation of three-parameter compensatory multidimensional item response models*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). *BILOG-MG 3.0* [Computer software and manual]. Lincolnwood, IL: Scientific Software International.

Footnotes

[1] Because TESTFACT only provides ability estimates for the primary factor when the bi-factor model is used, the item parameter file was re-configured to match the format of the item parameter file from a three-factor model, with *a*-parameters set to 0 for the secondary factor not measured by an item, and the ability estimates were obtained in a second run using this parameter file. This was possible because the factors in the bi-factor model are orthogonal; TESTFACT estimates abilities only for orthogonal factors. As a check, scores for the primary factor were estimated along with the item parameters and compared to the estimates in the second run after re-configuring the item parameter file; the score estimates were identical and the estimated standard errors were nearly the same.

[2] To be accurate, CONQUEST (Wu, Adams, & Wilson, 1998) will estimate EAP and ML scores for multidimensional IRT models, but only for models that are an extension of the Rasch family of models -- models with constant discrimination parameters.

Table 1

Distribution of American History/Political Science Subscale Scores

	Mean	SD	Minimum	Maximum
TESTFACT Bi-factor Model	0.00	0.90	-2.76	2.20
NOHARM Bi-factor Model	-0.01	0.88	-2.68	2.18
NOHARM 2-Factor Model	0.00	0.91	-2.78	2.21
Augmented Subscale Method	0.01	0.90	-2.71	2.03
Independent Unidimensional Model	0.01	0.87	-2.52	1.91

Table 2

Distribution of Global Issues Subscale Scores

	Mean	SD	Minimum	Maximum
TESTFACT Bi-factor Model	0.00	0.87	-2.92	2.05
NOHARM Bi-factor Model	-0.01	0.86	-2.87	2.03
NOHARM 2-Factor Model	0.00	0.88	-2.91	2.04
Augmented Subscale Method	0.01	0.87	-2.78	1.84
Independent Unidimensional Model	0.01	0.82	-2.55	1.44

Table 3

Correlations among American History/Political Science Subscale Scores

	TESTFACT Bi-factor Model	NOHARM Bi-factor Model	NOHARM 2-Factor Model	Augmented Subscale Method
NOHARM Bi-factor Model	1.000			
NOHARM 2-Factor Model	0.995	0.995		
Augmented Subscale Method	0.997	0.996	0.997	
Independent Unidimensional Model	0.970	0.971	0.981	0.975

Table 4

Correlations among Global Issues Subscale Scores

	TESTFACT Bi-factor Model	NOHARM Bi-factor Model	NOHARM 2-Factor Model	Augmented Subscale Method
NOHARM Bi-factor Model	0.999			
NOHARM 2-Factor Model	0.991	0.989		
Augmented Subscale Method	0.994	0.991	0.996	
Independent Unidimensional Model	0.943	0.944	0.962	0.950

Table 5

Correlations among Course Grades and American History/Political Science Subscale Scores

	Course	
	History (N = 861)	Political Science (N = 319)
TESTFACT Bi-factor Model	.32	.36
NOHARM Bi-factor Model	.32	.36
NOHARM 2-Factor Model	.31	.36
Augmented Subscale Method	.32	.36
Independent Unidimensional Model	.29	.35

Table 6

Correlations among Course Grades and Global Issues Subscale Scores

	Course				
	anthropology (N = 293)	economics (N = 741)	geography (N = 315)	political sci. (N = 90)	sociology (N = 221)
TESTFACT Bi-factor Model	.37	.29	.33	.66	.32
NOHARM Bi-factor Model	.37	.29	.33	.66	.31
NOHARM 2-Factor Model	.38	.30	.32	.66	.31
Augmented Subscale Method	.38	.30	.32	.67	.31
Independent Unidimensional Model	.35	.28	.30	.65	.24

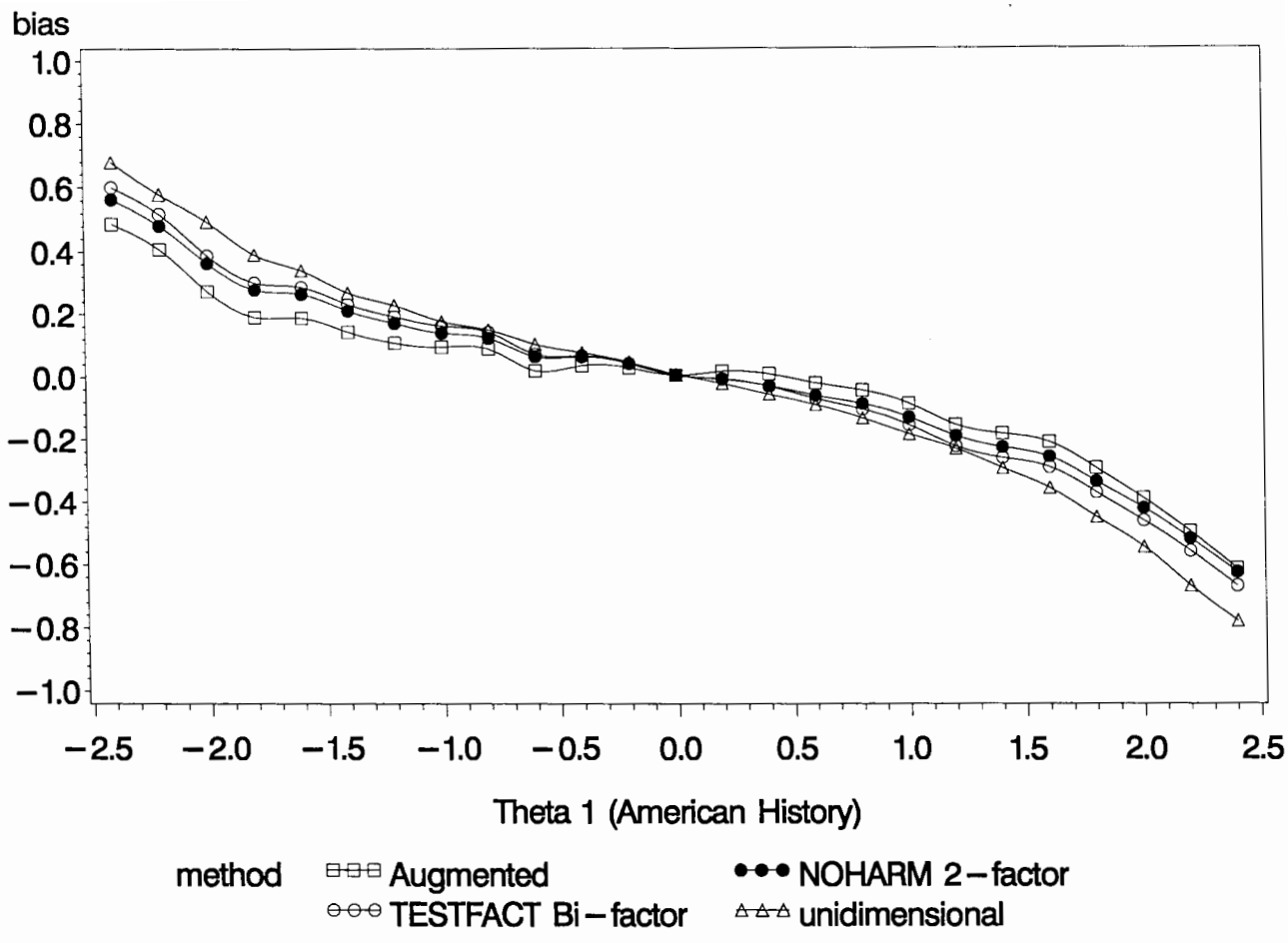
Figure Captions

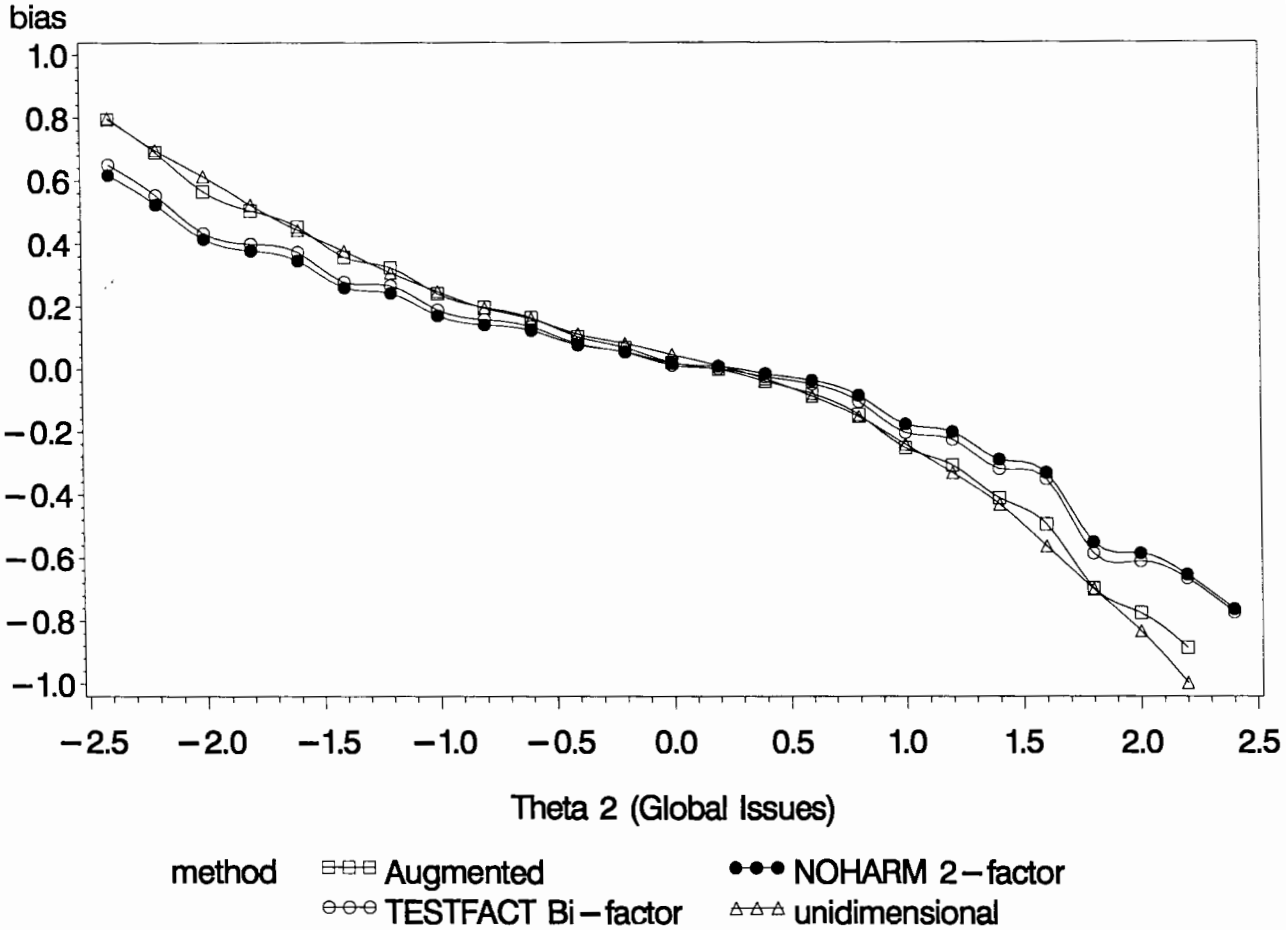
Figure 1. Bias of American History/Political Science Scores

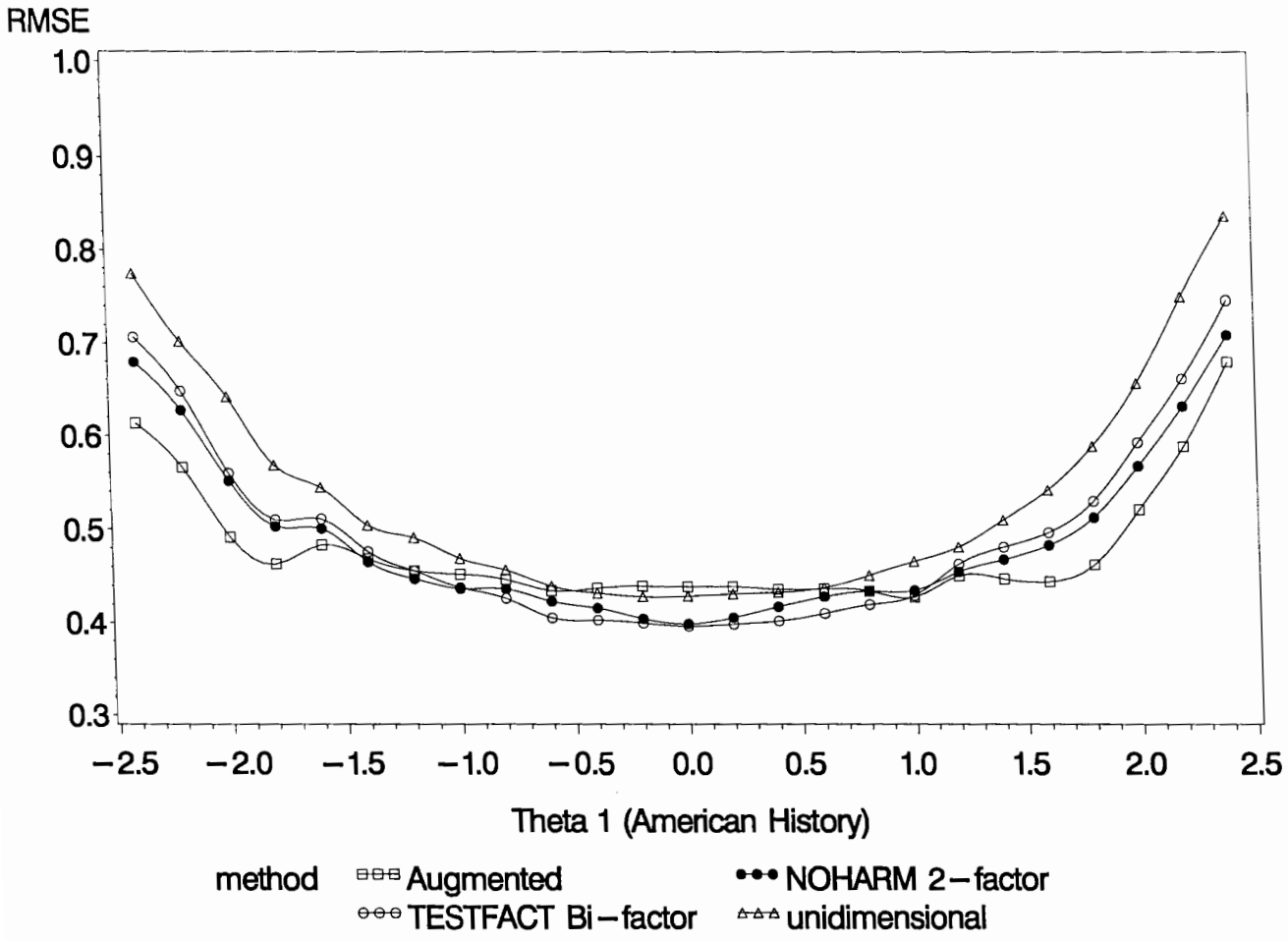
Figure 2. Bias of Global Issues Scores

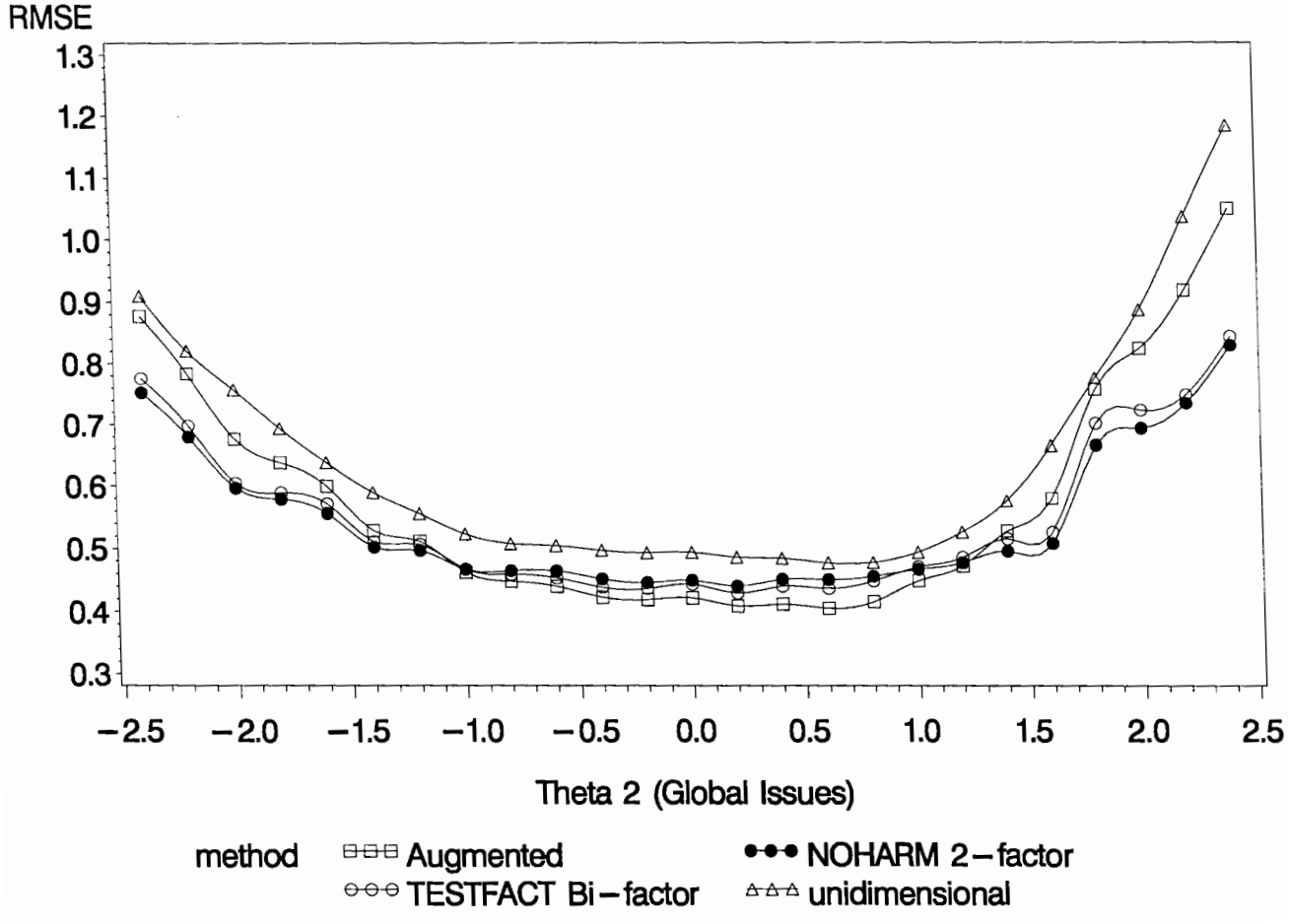
Figure 3. RMSE of American History/Political Science Scores

Figure 4. RMSE of Global Issues Scores


bias
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
−1.0
−2.5
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
Theta 1 (American History)
method
Augmented
NOHARM 2−factor
TESTFACT Bi−factor
unidimensional


bias
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
−1.0
−2.5
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
Theta 2 (Global Issues)
method
Augmented
NOHARM 2−factor
TESTFACT Bi−factor
unidimensional


RMSE
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
−2.5
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
Theta 1 (American History)
method
Augmented
NOHARM 2−factor
TESTFACT Bi−factor
unidimensional


RMSE
1.3
1.2
1.1
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
−2.5
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
Theta 2 (Global Issues)
method
Augmented
NOHARM 2−factor
TESTFACT Bi−factor
unidimensional