"Guessing" Parameter Estimates for Multidimensional IRT Models

Christine E. DeMars

James Madison University

(2005, April). Paper presented at the annual meeting of the American Educational Research Association, Montreal.

Abstract

Two software packages in common use for multidimensional item response theory (IRT) models require the user to input values for the lower asymptotes of the item response functions. One way of selecting these values is to estimate lower asymptotes with a one-dimensional IRT model and use those estimates as fixed values in the multidimensional model. This procedure was compared to simply setting the asymptotes to a reasonable value. The fixed value of the asymptote was varied to be somewhat higher or lower than the true value. For two-factor tests, the use of unidimensional asymptotes worked well, yielding correlations with true difficulty and discrimination parameters and root mean square error (RMSE) between estimated and true item response surfaces nearly comparable to setting the lower asymptotes to the true values. Setting the asymptotes too low produced better results than setting them too high. With four-factor tests, in contrast, the correlations were lower and the RMSEs were higher when the lower asymptotes were estimated through a unidimensional model. The estimates of the lower asymptotes from the unidimensional model tended to be too high, which likely caused the decreased accuracy of this procedure.

'Guessing' Parameter Estimates for Input to NOHARM and TESTFACT

TESTFACT (Bock, Gibbons, Schilling, Muraki, Wilson, & Wood, R., 2003) and NOHARM (Fraser, 1988) are software programs to estimate multidimensional item response theory (MIRT) models. TESTFACT uses full-information marginal maximum likelihood techniques to estimate the item parameters (Bock, Gibbons, & Muraki, 1988; Gibbons & Hedeker, 1992; Muraki & Engelhard, 1985), and NOHARM uses a polynomial approximation procedure (McDonald, 1997; 1999). Both packages apply the multidimensional normal ogive model:

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

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.

The vector $\mathbf{a}_i$ indicates how discriminating the item is along each dimension. Unless an item measures only a single dimension, the direction (in the appropriate dimensional space) of steepest slope will lie somewhere between the dimensions. Reckase (1985; 1997; Reckase & McKinley, 1991) defined the direction of steepest slope as

$$\cos\alpha_{ik} = \frac{a_{ik}}{\sqrt{\left(\sum a_{ik}^2\right)}}, \quad (2)$$

where α_{ik} is the angle with axis k for item i, and a_{ik} is the kth element of $\mathbf{a}_i$, the discrimination vector for item i. Reckase proposed that a generalized discrimination parameter could be:

$$\text{MDISC} = \sqrt{\left(\sum a_{ik}^2\right)}, \quad (3)$$

The lower asymptote, or guessing parameter, must be supplied to either package; it is not estimated with the other parameters. A common approach is to estimate the *c*-parameters using a unidimensional model and software appropriate for unidimensional models, then use these estimates in the multidimensional model (Jodoin & Davey, 2003; McLeod, Swygert, & Thissen, 2001; Miller & Hirsch, 1992; Zhang & Stone, 2004a). In the TESTFACT manual, it is noted that the *c*-parameters should not depend on the dimensionality of the model (p. 585). However, one potential problem is that while the true *c*-parameters may not depend on the dimensionality, the *estimation* of the *c*-parameters may depend on the accuracy of the estimation of the other parameters, which in turn may depend on the dimensionality of the model. Reciprocally, the accuracy of the estimation of the other parameters in the multidimensional model may depend on the accuracy of the *c*-parameter. For example, if one fixes the *a* or *c* in a unidimensional model, the estimated value of the free parameter will change to compensate for the fixed parameter. If *c* is fixed to zero, the estimate of the *a* will be less steep to better accommodate the data at the low end of the theta range, especially for difficult items (Yen, 1981). Li and Lissitz (2004) discussed how poorly estimated *c*-parameters can lead to large standard errors for the estimates of the *b*-parameters in the unidimensional 3PL model. In another context, Wainer and Wang (2000) found that when local dependencies among items in the same testlet were ignored, *c*-parameters were overestimated and *a*-parameters were overestimated on one test and underestimated on another. Because local-dependencies may be considered a form of multidimensionality where the secondary dimensions are nuisance dimensions, this problem may extend to other multidimensional models. If a multidimensional model is incorrectly specified as a unidimensional model, it could have an impact on the estimation of all of the parameters.

Zhang and Stone (2004a) compared the recovery of multidimensional item parameters and response functions by TESTFACT and NOHARM using c-parameters estimated through a unidimensional model using MULTILOG (indirect estimation) with the recovery by programs which estimated the c-parameters directly using the multidimensional model, WinBUGS (Spiegelhalter, Thomas, Best, & Lunn, 2003) and ASSEST (Zhang & Lu, 2001, cited in Zhang & Stone, 2004a). The difficulty and discrimination parameters were estimated somewhat better when the *c*-parameters were directly estimated than when they were estimated based on a unidimensional model. When comparing item response functions (IRFs), though, the RMSEs for TESTFACT and NOHARM were larger than one of the direct methods and smaller than the other.

A simpler approach would be to fix the *c*-parameters to a set value (as in Zhang & Stone, 2004b, for example). This could produce inaccuracies in the estimation of the other parameters, but these inaccuracies may be no worse than the inaccuracies produced by using estimates of the *c*-parameters from a unidimensional model. In the proposed study, data were simulated to compare the recovery of the item parameters and item response functions with fixed lower asymptotes (not necessarily fixed to the correct values) compared to lower asymptotes estimated through a unidimensional model.

Method and Data Source

Data were simulated to follow the multidimensional model given in Equation 1. To make the simulation easier, the equivalent logistic form of the model was used. Two and four factor orthogonal models were included. There were 40 items on each of 100 test forms. All lower asymptotes were set to .2. For the two dimensional models, MDISC values were drawn from a log-normal distribution with a mean of zero and standard deviation of 0.5. Discrimination values

were then calculated from the MDISC values such that the first 10 items had an angle of 15^{O} with the first factor axis and 75^{O} with the second factor axis, items 11-20 had angles of 30^{O} and 60^{O} with the first and second factors, items 21-30 had angles of 45^{O} with each factor, items 31-35 had angles of 60^{O} and 30^{O} with the first and second factors, and items 36-40 had angles of 75^{O} and 15^{O} respectively. With this scheme, each factor had some items that measured it more than the other factor, but more items were more heavily in the direction of factor 1 to reduce the indeterminacy of which factor would be likely to be the "first" factor. For the four dimensional models, each item loaded on only two factors. Eighteen items loaded primarily on the first factor, 12 items primarily on the second, 7 primarily on the third, and 3 primarily on the fourth. The direction of measurement for each item was 15^{O} from the primary factor axis and 75^{O} from one of the other factors. Again, this design was chosen to minimize the indeterminancy in the factor order.

Difficulty values were generated as the product of MDISC and a random draw from a standard normal distribution. 100 test forms were simulated, with different sets of items on each form. Two or four uncorrelated θ's were drawn from standard normal distributions for each of 1000 simulees. Uncorrelated θ's were used to be consistent with the exploratory analysis; in the exploratory mode both NOHARM and TESTFACT extract uncorrelated factors initially. For each item score, the simulee's probability of correct response was calculated based on the model; if a random number from a uniform distribution between 0 and 1 was less than the model-based probability, the item was coded correct. Random numbers were selected using the RANNOR( ) or the RANUNI( ) function in SAS, with a time-of-day seed.

The discrimination and difficulty parameters were recovered using TESTFACT and NOHARM, with the following values for the lower-asymptotes: .10, .15, .20, .25, .30 or

estimated from a unidimensional model in BILOG-MG. Accuracy of the recovered item parameters was assessed by correlations with the true parameters, mean difference between the estimated and true parameters, standard deviation of the difference between the estimated and true parameters, and the RMSE between the estimated and true item functions.

Results

Two-Dimensional Tests

Fixing *c* to .3 led to estimation problems in TESTFACT. For 10 of the 100 test forms the iterations stopped and an error message was displayed. For another 41 of the forms, either the order of the dimensions was switched or one or both of the dimensions ended up reverse-coded (high θ's indicated low raw scores). The .3 condition was also problematic in NOHARM; for 84 of the 100 tests NOHARM automatically changed some of the *c*'s from .3 to 0, presumably because the proportion of simulees getting the items correct was too low for a *c* of .3. The frequency of these problems is noted in Table 1. Because these problems occurred much more frequently in this condition, the *c* fixed to .3 condition was dropped from further analyses. In the other conditions, in the rare cases where the iterations stopped and an error message was displayed, the starting values were adjusted such that the discrimination parameters above 2.5 were set to 2.5 and difficulty parameters below -3 were set to -3. If the metrics were in the reverse of the intended direction, the discriminations were multiplied by -1. If the dimensions were switched they were interchanged for further analyses. This solution was somewhat awkward; if the item parameters had been selected to measure both dimensions equally, designation of one factor as "factor 1" would have been arbitrary; however, the item parameters, as described above, were selected such that the test as a whole would be weighted toward one factor so it seemed somewhat problematic when the other factor emerged as the first factor.

Within each test form, the correlations between the true (generating) item parameters and the corresponding estimated item parameters were calculated. The average correlations are shown in Table 2; correlations were transformed using Fisher's *r*-to-*z* transformation before averaging. The correlation between the true and estimated item difficulty was nearly one in all conditions, except for NOHARM when the lower-asymptote was set too high (.25). For TESTFACT, the correlation between the true and estimated discriminations was highest when the lower-asymptote was set to the true value (.2) but was only slightly lower when the lower asymptote was increased or decreased by .05 (set to .15 or .25). The correlation decreased to .90 when the asymptote was set to .1. In NOHARM, the correlations between the true and estimated discriminations were somewhat lower than they were in TESTFACT, particularly when the asymptote was set too high. Setting the asymptote too high yielded unacceptable results in NOHARM. In fact, this condition even noticeably affected the estimation of the item difficulty.

In interpreting these correlations, keep in mind that the true asymptotes were equal for all items, and in all of the conditions with a fixed lower asymptotes the asymptotes were "off" to the same degree for all items. If they had been set too high for some items and too low for other items, the correlations would likely have been lower. This may be why the correlations were lower in NOHARM when *c* was fixed too high; in this condition, as described earlier sometimes the *c* was automatically set to 0 (too low) for some items while it was left too high for the remaining items. But in TESTFACT and most of the time in NOHARM the fixed *c* was equally accurate or inaccurate for all items. Any consistent effects of misspecifying the asymptotes do not appear in the correlations but can instead be seen by examining the mean difference between the estimated and true parameters.

The mean difference between the estimated and true parameters is shown in Table 3, with the standard deviation of the difference between the estimated and true parameters in parentheses. This mean difference does not reflect any bias toward the mean that might occur for extreme items due to TESTFACT's use of prior distributions; instead the mean difference shows whether there was an overall tendency for the items to be estimated to be too difficult/too easy/too discriminating/not discriminating enough. For a few conditions, the mean difference was larger for TESTFACT. When *c* was set too high, NOHARM was considerably less accurate than TESTFACT; the mean difference and the standard deviation of the differences was greater for NOHARM. In both packages, as would be expected, the items appeared easier[1] when the lower asymptote was set too low and harder when the lower asymptote was set too high. The discriminations were estimated to be less steep to compensate for lower asymptotes set too low.

Perhaps more important than the recovery of the individual item parameters is the recovery of the item response surface (IRS), also called the item characteristic surface (ICS). To approximate the distance (in a direction parallel to the probability axis, perpendicular to the θ-plane) between the true and recovered IRS's throughout the ability space, the difference between the probability of correct response based on the true parameters and the probability based on the estimated parameters was estimated at 81 quadrature points (9 equally spaced quadrature points from -4 to 4 for each ability). These differences were squared and weighted based on weights from a multivariate standard normal distribution with independent dimensions. The square root of this weighted average was an approximation of the root mean square error (RMSE) of the IRS. Notice that these differences are taken only in one direction in space and do not represent

[1] In Equation 1, d_i was defined such that easier items had more positive values. Thus, positive bias indicates the estimated d was higher, or easier, than the true d.

the volume between the surfaces. These RMSEs, averaged across the 2000 items, are displayed in Table 4.

The RMSE was about 1/3 larger for NOHARM than for TESTFACT in each condition. As was true for the correlations between true and estimated discriminations, setting the asymptotes too high led to larger RMSE than setting them too low, especially for NOHARM. Estimating the asymptotes from BILOG yielded results similar to setting the asymptotes to the true value of .2 or to .15.

Four-Dimensional Tests

For the four-dimension tests, the most common problem in TESTFACT (see Table 5) was that the third or fourth dimension was frequently reverse-coded (high θ's indicated low raw scores), which was addressed by multiplying the affected discrimination parameters by -1. In NOHARM, the 2nd, 3rd, and 4th factors did not seem to be recovered well. It was not that the order of the factors was switched as occasionally happened in TESTFACT; the factors that emerged did not appear to correspond with any of the original factors. The varimax-rotated factors were no better matched to the original factors than the unrotated factors. Different starting values were tried: instead of the default initial slopes of 0.5, initial slopes of 0.8 for the factor which the item measured most heavily, 0.3 for the other factor the item measured and 0 for the remaining two factors. This did not work well; the estimation routine went through few, often only one, iterations and the final slopes were virtually unchanged from the starting values. Another set of starting values, 0.5 for the primary factor that the item measured and 0.2 for the other three factors, was attempted, but again the final values for the slopes were only slightly changed from the initial values. An attempt was made to anchor the solution by fixing one item to have zero loadings on the third and fourth factors, another to have zero loadings on the 2nd and

4th factors, and another to have zero loadings on the second and third factors. This did not improve the correlations for factors 2 through 4 very much. Because none of these efforts helped much, the results discussed below are those from using the default starting values and a completely exploratory model. The sign of the discriminations were changed for any factors for which the discriminations had a negative correlation with the generating discriminations, just as they were for the "reversed" factors in TESTFACT even though the term "reversed" is not quite appropriate here because the correlations were often so low that it was not really the same factor at all.

The correlations between the true and estimated parameters are shown in Table 6. For the item difficulties, results were similar for TESTFACT and NOHARM. These correlations were somewhat lower when *c* was estimated in BILOG or set above the true value. Overestimation of the lower asymptote in BILOG seemed to be responsible for these effects. The average estimated asymptote was .25, .27, .31, and .32 for items which loaded primarily on factor 1, factor 2, factor 3, or factor 4 respectively. Discriminations for later factors had lower correlations with their true factors; this decrease was small for TESTFACT but very large for NOHARM.

Because of the difficulties in recovering the intended factors in NOHARM, NOHARM was also run in a confirmatory mode, with the factor loadings free only for the factors a given item was generated to load on. The correlations among the factors were freed instead of fixed to zero because of problems running the software when these correlations were fixed to zero. These estimated correlations were nearly unbiased, averaging zero when the lower asymptotes were set to .1, .15, or .2, averaging 0.01 when the lower asymptotes were set to .25, and averaging 0.02 when the lower asymptotes were estimated in BILOG. The correlations between the estimated and true parameters, not including the parameters fixed to zero, are shown at the bottom of Table

6. For factors 2-4, these correlations are much higher than the exploratory correlations from NOHARM and are fairly comparable to the TESTFACT correlations. Of course, when making comparisons to TESTFACT one needs to keep in mind that the TESTFACT results were from an exploratory model.

For the individual item parameters, the mean difference between the estimated and true parameters, with the standard deviation of the difference in parentheses, is shown in Table 7. For NOHARM, the results are shown only for the confirmatory model because the exploratory model did not seem to be recovering the same factors. For each item, two of the four true discrimination parameters were zero. For NOHARM, these parameters were fixed to zero and were not included in the calculation of the mean difference. For TESTFACT, the differences for these discrimination parameters was calculated separately from those for the non-zero parameters. These discriminations generally averaged close to zero, but the standard deviations are large enough to indicate that for some items or in some replications the estimated discriminations were meaningfully different from zero. The standard deviations, however, were still lower than those for the other discriminations.

As with the two-dimensional tests, the items were estimated to be easier when the lower asymptotes were set too low and more difficult when the asymptotes were set too high. Because the asymptotes estimated from the unidimensional model tended to be too high, the difficulties were considerably biased when these asymptotes were used. The discriminations were on average slightly too low, somewhat more so in TESTFACT, when the asymptotes were set to the true value of .2. Using asymptotes that were too high, either from the unidimensional model or set to .25, led to a positive bias in the discriminations, especially for NOHARM. Using asymptotes that were too low led to the opposite effects; the discriminations were too low. The

standard deviations of the differences were generally somewhat higher for NOHARM. Also, the standard deviations were highest when the asymptotes were estimated with a unidimensional model.

The RMSE of the IRS's, estimated on 6561 quadrature points (9 for each ability), are displayed in Table 8. The RMSE from the NOHARM confirmatory model was consistently somewhat smaller than the RMSE from the TESTFACT exploratory model; again it is important to note that not only the software but the exploratory/confirmatory mode differs. The RMSE was larger when the lower asymptotes were estimated from a unidimensional model or when they were set too high. This would be expected from the results in Tables 6 and 7.

Discussion and Conclusions

Fixed vs. Estimated Lower Asymptote

When there were only two ability factors, estimating the lower asymptotes through a unidimensional model was effective. This approach for estimating the lower asymptotes yielded a mean difference and standard deviation of the differences between the true and estimated item parameters and RMSE of the IRS comparable to the results from a model with the lower asymptotes fixed at the true value of .2. Setting the lower asymptote to .15 instead of .20 had little effect on the RMSE of the IRS, but setting it lower to .1 or setting it even a small amount too high to .25 had a bigger effect on the RMSE of the IRS. Setting the asymptote too high was particularly problematic in NOHARM. Based on these results, it seems better to estimate the lower asymptote from a unidimensional model than to set it to a reasonable value which might be somewhat too high or too low. This might be particularly true for a test where the lower asymptotes varied from item to item; the chosen asymptote would be too high for some items and too low for others, leading to bias in different directions for different items.

However when there were four factors, the lower asymptotes tended to be overestimated in the unidimensional model. This led to mean differences, standard deviation of differences, and RMSE similar to fixing the asymptote too high. With four factors, less error would likely be introduced by setting the asymptotes to a value at or somewhat lower than chance guessing rather than estimating them through a unidimensional model. Zhang and Stone's (2004a) approach of estimating the lower asymptotes directly in the multidimensional model, though considerably more complex, should also be considered in this context (see Appendix).

Poor Recovery of Higher Dimensions in NOHARM in Exploratory Model

When there were only two dimensions, NOHARM was only marginally less effective than TESTFACT at recovering the item parameters and the IRS. However, when there were four dimensions the second, third, and fourth factors had only small correlations with the generating factors. Other studies (Béguin & Glas, 2001; Gosz &Walker, 2002; Knol & Berger, 1991, Miller, 1991; Zhang & Stone, 2004a) have found that NOHARM and TESTFACT performed relatively similarly in terms of recovering multidimensional parameters in an exploratory mode. However, most of these studies included only two factors and only Béguin and Glas used more than three. In the present study, while NOHARM had consistently lower correlations and higher RMSE than TESTFACT even when there were only two factors, the differences were not that large except when the lower asymptote was set too high. The performance differences between the software packages were much greater when there were four factors, a condition not explored in most of the reviewed studies. Finch and Habing's (2004) work hints at some problems in extracting larger numbers of factors in NOHARM. They obtained unexpected results when they attempted to recover more factors than existed in the data. Their data were generated using two factors, but the items that should have loaded on the second factor loaded on a later factor when three or

more factors were extracted. In the present study though, it was not clear or consistent that items from the second factor loaded on a later factor; more frequently after the first factor none of the extracted factors seemed to have much relationship to any of the true factors. Also, it is important to note that Béguin and Glas extracted five factors in an exploratory analysis and the median absolute residuals of the discriminations were only noticeably larger in NOHARM for one of the factors. There may have been something about the particular factor structure used for the present study that created difficulties for NOHARM's procedures but not for TESTFACT's.

Implications

This study is useful for practitioners who are considering estimating lower asymptotes with a one-dimensional IRT model and using those estimates as fixed values in a multi-dimensional model. For two factors, this procedure worked well, yielding correlations with true parameters and RMSE between estimated and true IRS nearly comparable to setting the lower asymptotes to the true values (which of course would not be known in actual practice). With four factors, the correlations were lower and the RMSEs were higher when the lower asymptotes were estimated through a one-dimensional model, probably because the lower asymptotes tended to be overestimated. With four or more factors, then, it might be better to set the lower asymptotes to a fixed value, perhaps a bit lower than the reciprocal of the number of response options.

References

Béguin, A. A., & Glas, C. A. W. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika, 66*, 541-562.

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.

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

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

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.

Jodoin, M. G., & Davey, T. (2003, April). *A multidimensional simulation approach to investigate the robustness of IRT common item equating*. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Li, Y. H., & Lissitz, R. W. (2004). Applications of the analytically derived asymptotic standard errors of item response theory item parameter estimates. *Journal of Educational Measurement, 41*, 85-117.

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., & Hirsch, T. M. (1992). Cluster analysis of angular data in applications of multidimensional item-response theory *Applied Measurement in Education, 5*, 193-211.

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

Spiegelhalter, D., Thomas, A., Best, N., & Lunn, D. (2003). WinBUGS 1.4 [computer software]. Cambridge, UK: MRC Biostatistics Unit, Institute of Public Health.

Wainer, H., & Wang, X. (2000). Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement*, *37*, 203-220.

Zhang, B., & Stone, C. (2004a, 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.

Zhang, B., & Stone, C. (2004b, April). *Type I error rates and empirical power of an item fit index based on total scores for multidimensional item response models*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5*, 245-262.

Table 1

Problems Encountered in Running the 2-Dimensional Models

	TESTFACT			NOHARM	
	error	dimensions switched	reverse-coded	dimensions switched	c's adjusted
c input from BILOG	0	3	0	0	0
c set to .2	0	1	0	1	0
c set to .1	0	4	0	1	0
c set to .15	0	0	0	1	0
c set to .25	2	18	1	1	30
c set to .3	10	32	9	1	84

Table 2

Correlations between True and Recovered Parameters, 2-Dimensional Model

	TESTFACT			NOHARM		
	d	a_1	a_2	d	a_1	a_2
c input from BILOG	.99	.95	.95	.98	.91	.92
c set to .2	.99	.97	.97	.99	.92	.93
c set to .1	.98	.90	.90	.98	.86	.88
c set to .15	.99	.95	.95	.98	.89	.91
c set to .25	.99	.95	.95	.87	.69	.79

Table 3

Mean Difference between Estimated and True Parameters, 2-Dimensional Model

	TESTFACT			NOHARM		
	b	a_1	a_2	b	a_1	a_2
c input from BILOG	-0.07 (0.19)	-0.06 (0.13)	0.00 (0.12)	-0.01 (0.22)	0.02 (0.20)	-0.02 (0.19)
c set to .2	-0.06 (0.13)	-0.06 (0.11)	0.00 (0.10)	-0.01 (0.20)	0.02 (0.19)	-0.01 (0.17)
c set to .1	0.26 (0.32)	-0.19 (0.18)	-0.12 (0.17)	0.25 (0.35)	-0.19 (0.22)	-0.15 (0.19)
c set to .15	0.13 (0.24)	-0.14 (0.14)	-0.07 (0.13)	0.16 (0.29)	-0.12 (0.20)	-0.09 (0.17)
c set to .25	-0.26 (0.22)	0.00 (0.15)	0.09 (0.14)	-0.65 (1.76)	0.54 (1.07)	0.24 (0.71)

Note. standard deviation of the difference between the estimated and true parameters is in parentheses.

Table 4

RMSE of the Item Response Surface, 2-Dimensional Model

Condition	TESTFACT RMSE	NOHARM RMSE
c input from BILOG	.028	.037
c set to .2 (true value)	.027	.036
c set to .1	.035	.042
c set to .15	.028	.037
c set to .25	.037	.058

Table 5

Problems Encountered in Running the 4-Dimensional Models

	TESTFACT			NOHARM	
	error	dimensions switched	one or more factors reverse-coded	some factors did not match generating factors	c's adjusted
c input from BILOG	0	2	91	100	0
c set to .2	0	1	95	100	0
c set to .1	0	3	94	100	0
c set to .15	0	1	96	100	0
c set to .25	1	19	71	100	35

Table 6

Correlations between True and Recovered Parameters, 4-Dimensional Model

	d	a_1	a_2	a_3	a_4
TESTFACT					
c input from BILOG	.93	.96	.96	.95	.89
c set to .2	.99	.99	.98	.97	.93
c set to .1	.98	.96	.96	.96	.93
c set to .15	.99	.98	.98	.97	.94
c set to .25	.99	.97	.97	.97	.87
NOHARM, Exploratory					
c input from BILOG	.94	.87	.41	.21	.15
c set to .2	.98	.88	.34	.22	.15
c set to .1	.97	.86	.35	.23	.15
c set to .15	.98	.87	.35	.22	.16
c set to .25	.96	.88	.30	.21	.16
NOHARM, Confirmatory					
c input from BILOG	.92	.90	.90	.91	.89
c set to .2	.99	.96	.96	.96	.93
c set to .1	.98	.90	.93	.94	.92
c set to .15	.99	.93	.94	.95	.93
c set to .25	.96	.91	.93	.94	.92

Table 7

Mean Difference between Estimated and True Parameters, 4-Dimensional Model

	d	a_1	a_2	a_3	a_4
TESTFACT (True parameters ≠ 0)					
c input from BILOG	-0.34 (0.62)	0.00 (0.26)	0.06 (0.38)	0.11 (0.45)	0.07 (0.51)
c set to .2	-0.06 (0.19)	-0.05 (0.12)	-0.05 (0.12)	-0.05 (0.16)	-0.05 (0.25)
c set to .1	0.26 (0.34)	-0.21 (0.22)	-0.17 (0.21)	-0.16 (0.18)	-0.13 (0.17)
c set to .15	0.12 (0.26)	-0.14 (0.17)	-0.12 (0.16)	-0.11 (0.15)	-0.10 (0.15)
c set to .25	-0.30 (0.30)	0.02 (0.23)	0.03 (0.19)	0.04 (0.20)	-0.01 (0.41)
TESTFACT (True parameters = 0)					
c input from BILOG		0.01 (0.10)	0.01 (0.09)	0.00 (0.10)	-0.03 (0.16)
c set to .2		0.02 (0.06)	0.01 (0.06)	0.00 (0.07)	-0.04 (0.10)
c set to .1		0.01 (0.04)	0.01 (0.05)	0.00 (0.05)	-0.03 (0.07)
c set to .15		0.01 (0.05)	0.00 (0.05)	0.00 (0.06)	-0.03 (0.07)
c set to .25		-0.02 (0.15)	0.00 (0.10)	0.02 (0.12)	-0.02 (0.28)
NOHARM (Confirmatory Model)					
c input from BILOG	-0.32 (0.82)	0.16 (0.50)	0.20 (0.60)	0.21 (0.51)	0.09 (0.30)
c set to .2	0.00 (0.19)	-0.01 (0.15)	-0.02 (0.15)	-0.03 (0.14)	-0.03 (0.15)
c set to .1	0.25 (0.38)	-0.21 (0.24)	-0.17 (0.22)	-0.15 (0.19)	-0.10 (0.18)
c set to .15	0.15 (0.31)	-0.14 (0.20)	-0.12 (0.19)	-0.11 (0.17)	-0.07 (0.16)
c set to .25	-0.29 (0.54)	0.23 (0.40)	0.16 (0.33)	0.10 (0.28)	0.04 (0.21)

Note. standard deviation of the difference between the estimated and true parameters is in parentheses.

Table 8

RMSE of the Item Response Surface, 4-Dimensional Model

Condition	TESTFACT	NOHARM (confirmatory)
c input from BILOG	.063	.053
c set to .2	.038	.027
c set to .1	.042	.037
c set to .15	.037	.031
c set to .25	.067	.039

Appendix

Direct Estimation of Lower Asymptotes

Zhang and Stone (2004a) found that using the MCMC method to estimate the lower asymptotes along with the difficulties and discriminations yielded more accurate estimates of the item parameters as well as the item response function. To check how well this method worked with the data in the present study, a small simulation study was conducted using 10 of the replications from the four-dimensional data set.

The model was specified in WinBUGS (Spiegelhalter, Thomas, Best, & Lunn, 2003) in its logistic form. The priors for θ were multivariate standard normal, with zero correlations among the θ's. The priors for *a*'s were log-normal with log-mean of -0.695 and variance of 4 (precision of .25), constrained to the interval [0, 3]. The log-mean of -0.695 corresponded to a mean 0.5, which is the default starting value in NOHARM. A variance of 4 was diffuse enough not to influence the estimated values very much unless they were very extreme. The priors for *d*'s were normal (0,4), constrained to the interval [-4, 4]. Again, a relatively diffuse prior was used so that only the most extreme estimates would be affected by the prior. The priors for *c*'s were beta (5,17), constrained to the interval [0, .3]. This beta distribution is the default in BILOG and corresponds to a mean of 0.2. Starting values were 0 for θ, .5 for *a*, 0 for *b*, and 0 for *c*. A single chain was run for each replication; the first 1000 iterations were treated as the burn-in; examination of the trace plots showed that parameters seemed to converge well before the 1000th iteration. Parameter estimates were based on the mean of the next 5000 iterations.

The average *c* was 0.21, with a standard deviation of 0.02, so the *c*'s were estimated well and not overestimated as they were when using the unidimensional model. As with TESTFACT, the factors were often "switched" and were re-ordered before calculating correlations and

differences between estimated and true values. Table A1 has results parallel to those in Tables 6 and 7 in the main text. The correlations with the true parameters were comparable to the correlations obtained in the ideal condition (where the lower asymptotes were set to the true value) in TESTFACT and the confirmatory NOHARM. The same was true of the mean difference for the parameters which did not have values of zero, except that the mean difference for the discriminations was slightly positive using MCMC but slightly negative in TESTFACT and NOHARM. The mean difference from the MCMC procedure was greater for the discrimination parameters with values of zero; this may have been an artifact of using log-normal priors on the discrimination parameters. With a log-normal distribution, the estimates can never be less than zero, so if there is any variance at all the average of the estimates will be greater than zero. A normal distribution might have yielded estimates centered closer to zero. The RMSE for the item response surface was 0.034, again comparable to the results from setting the lower asymptote to the true value of 0.2 in TESTFACT and NOHARM.

With four dimensions, then, where estimating the lower asymptotes through a unidimensional model led to considerably less accurate estimates of the difficulty and discrimination parameters, the MCMC procedure would be preferred in terms of accuracy. However, MCMC takes considerably more computer time (each replication for this study took approximately 22 hours on a 1.8 Ghz Pentium 4) and considerably more user knowledge then TESTFACT or NOHARM. WinBUGS has made it possible for users to run MCMC without programming their own MCMC routines. However, the program is a general-purpose program, not designed specifically for IRT. It allows maximum flexibility in specifying the model and priors; the downside is this means greater opportunity for user misspecification. Many

practitioners may be better off using TESTFACT or NOHARM and simply specifying reasonable values for the lower asymptotes, as suggested in the Implications section.

Table A1

Correlations between True and Recovered Parameters, 4-Dimensional Model

	d	a_1	a_2	a_3	a_4
Correlations	.99	.98	.98	.97	.93
Mean Difference	-0.04	0.03	0.05	0.04	0.06
(True parameters ≠ 0)	(0.19)	(0.14)	(0.14)	(0.15)	(0.21)
Mean Difference		0.06	0.05	0.06	0.08
(True parameters = 0)		(0.03)	(0.03)	(0.04)	(0.05)

Note. standard deviation of the difference between the estimated and true parameters is in parentheses.