Online Supplemental Material Document for "A More Flexible Bayesian Multilevel Bifactor Item Response Theory Model"

Ken A. Fujimoto

Loyola University Chicago

Author Note

Ken A. Fujimoto, Program in Research Methodology, School of Education, Loyola University Chicago.

Online Supplemental Material Document for "A More Flexible Bayesian Multilevel Bifactor Item Response Theory Model"

This online supplemental material includes appendices that go with the main article. All of the notations, indices, and acronyms used in the main article are maintained here.

Online Appendix A

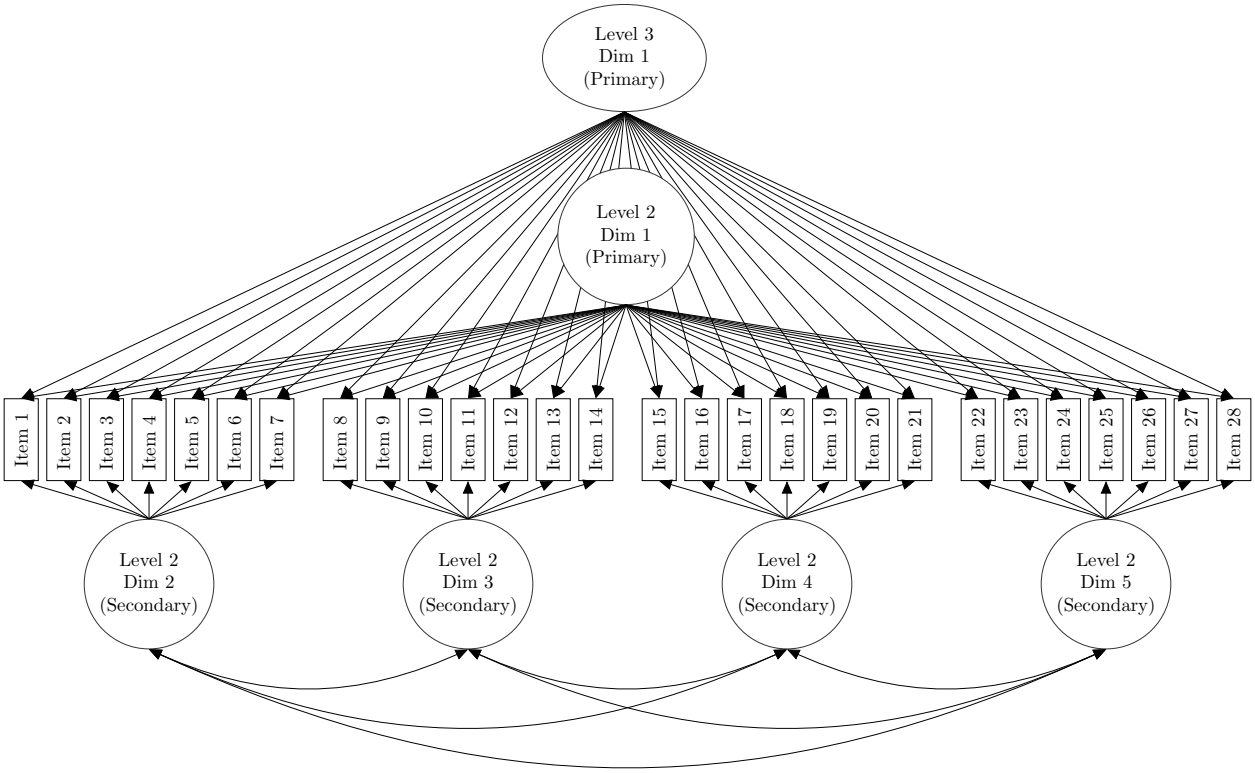
This appendix includes a summary of the nine models used in the main article (see Table A1) and visualizations of the dimensional structures with which these models were specified (see Figure A1). Figures A1a and A1b are visualizations of the dimensional structures specified in the models used to generate the data for the simulations.

Table A1
The Characteristics of the Nine Models Used in the Main Article

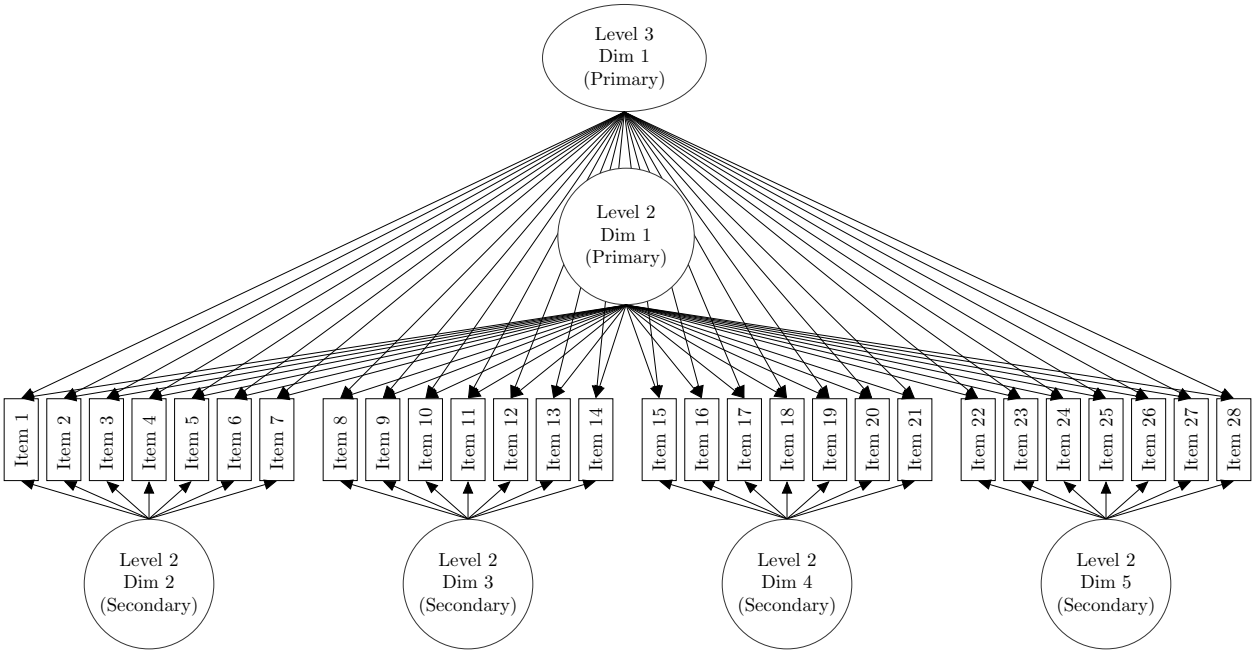
	Level 3		Level 2			Dimensional
	$\tilde{\alpha}_k$	$\sigma_{\tilde{\theta}1}$	Σ	$\rho_{dd'}$	ICC_{k1}	Structure
Three Levels: Unconstrained Discriminations						
Model 1: Correlated[a]	$\tilde{\alpha}_{k1} \neq \alpha_{k1}$	1	Σ	Free	Varying	Subfigure (a)
Model 2: Orthogonal[b]	$\tilde{\alpha}_{k1} \neq \alpha_{k1}$	1	Identity	0	Varying	Subfigure (b)
Model 3: Unidimensional[c]	$\tilde{\alpha}_{k1} \neq \alpha_{k1}$	1	1	None	Varying	Subfigure (c)
Three Levels: Constrained Discriminations						
Model 4: Correlated[a]	$\tilde{\alpha}_{k1} = \alpha_{k1}$	Free	Σ	Free	Constant	Subfigure (a)
Model 5: Orthogonal[b]	$\tilde{\alpha}_{k1} = \alpha_{k1}$	Free	Identity	0	Constant	Subfigure (b)
Model 6: Unidimensional[c]	$\tilde{\alpha}_{k1} = \alpha_{k1}$	Free	1	None	Constant	Subfigure (c)
Two Levels: Ignores Cluster Effects						
Model 7: Correlated[a]	$\tilde{\alpha}_{k1} = 0$	0	Σ	Free	None	Subfigure (a)
Model 8: Orthogonal[b]	$\tilde{\alpha}_{k1} = 0$	0	Identity	0	None	Subfigure (b)
Model 9: Unidimensional[c]	$\tilde{\alpha}_{k1} = 0$	0	1	None	None	Subfigure (c)

Note. "Identity" indicates a $D \times D$ matrix in which the elements of the main diagonal are fixed to 1 and all off-diagonal elements are 0. $\rho_{dd'}$ is the correlation between dimensions d and d', where $d \neq d'$ and both of these indices are greater than 1 because the first dimension represents the primary dimension, which is specified to be orthogonal to the secondary dimensions. A value of 0 or 1 indicates that the parameter was fixed to that value. ICC_{k1} is the intraclass correlation for item k. The subfigures in the column with the heading "Dimensional Structure" go with the subfigures in Figure A1.
[a]Allows the secondary dimensions to correlate with each other.
[b]The secondary dimensions are specified to be orthogonal to each other.
[c]This model did not include any secondary dimensions, making $\boldsymbol{\alpha}_k$, $\boldsymbol{\theta}_i$, and Σ scalars (i.e., $\boldsymbol{\alpha}_k = \alpha_{k1}$, $\boldsymbol{\theta}_i = \theta_{i1}$, and $\Sigma = \sigma^2 = 1$).

(a) The Correlated Multilevel Bifactor Structure



(b) The Orthogonal Multilevel Bifactor Structure



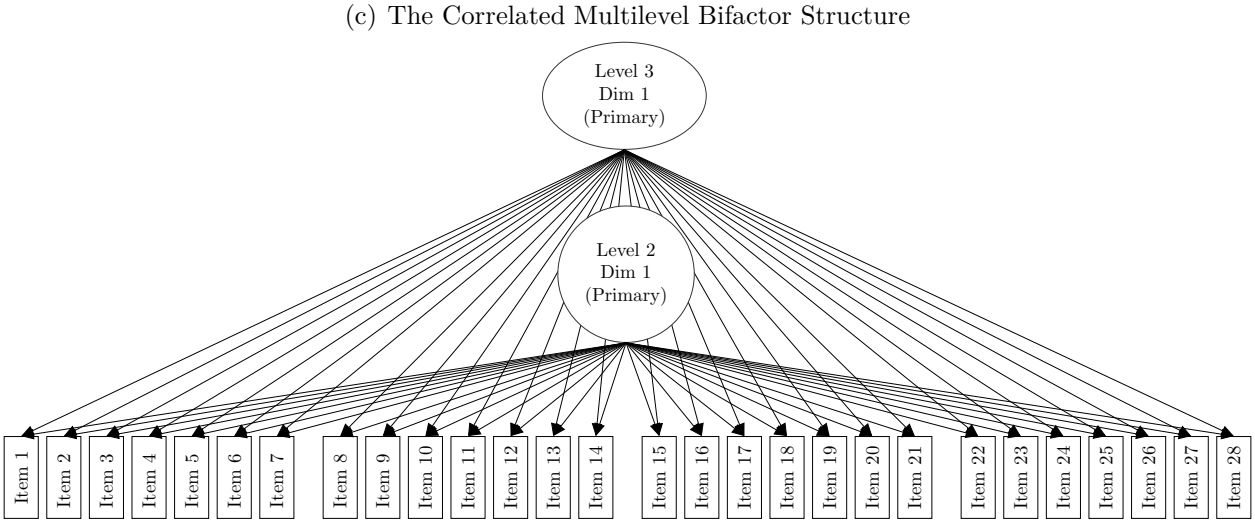


Figure A1. The following are visualizations of the dimensional structures specified in the models: (a) the multilevel bifactor structure with correlated secondary dimensions; (b) the multilevel bifactor structure with orthogonal secondary dimensions; and (c) the multilevel unidimensional structure.

In each subfigure, the oval is the Level 3 dimension, and the circle above the items is the Level 2 primary dimension. The circles below the items are the secondary dimensions (for the multilevel bifactor structures). Each straight line from a dimension to an item represents an item discrimination. The arced lines with arrows connecting secondary dimensions in subfigure (a) indicate that the secondary dimensions can be correlated with each other. "Dim" represents dimension.

For the unconstrained discrimination subclass of models, different sets of item discriminations were estimated for Levels 2 and 3. For the constrained discrimination subclass of models, the Level 3 item discriminations were constrained to be equal to their corresponding Level 2 discriminations (i.e., $\tilde{\alpha}_{k1} = \alpha_{k1}$ for all k). The two-level models can be viewed as three-level models but with the Level 3 item discriminations set to 0 (i.e., $\tilde{\alpha}_{k1} = 0$ for all k).

Online Appendix B

This appendix provides details on how the values for the parameters of the data generation models were obtained. The Level 2 dimensional positions ($\boldsymbol{\theta}_i$) were randomly drawn from $\mathcal{N}_5(\mathbf{0}, \Sigma_*)$, with Σ_* appropriately specified for the correlational condition (see the main article for these conditions). The Level 3 latent trait dimensional positions ($\tilde{\theta}_{j(\cdot)1}$) were randomly drawn from $\mathcal{N}(0, 1)$. Regarding the items, each one discriminated on three dimensions (the dimension at Level 3, the primary dimension at Level 2, and a secondary dimension also at Level 2). Thus, for each item, three values were randomly drawn from a uniform distribution ranging from 1 to 3, $U(1, 3)$. Then, the drawn values within a dimension were rescaled to overall scaling targets of 1.20, 1.00, 1.00, 1.00, and 1.00 for the first through the fifth dimension at Level 2, and an overall scaling target of 0.27 for the dimension at Level 3. The rescaling of the drawn values for the Level 2 discriminations proceeded as follows:

$$\alpha_{kd} = \alpha^*_{kd} \times \frac{\sigma^*_d}{\left(\prod_{k=1}^{K} \alpha^{*\mathbf{1}(0<\alpha^*_{kd})}_{kd}\right)^{1/K_d}},$$

where α_{kd} is the rescaled value used for item k's discrimination on dimension d in the data generation model, α^*_{kd} is the value drawn from $U(1, 3)$, σ^*_d is the target scaling factor for dimension d, $\mathbf{1}(0 < \alpha^*_{kd})$ is an indicator function that returns 1 when item k discriminates on dimension d and 0 otherwise, and K_d is the number of items discriminating on dimension d. The values drawn for the Level 3 discriminations were rescaled in the same way as the Level 2 discrimination values.

The rescaling of the discrimination values was performed for a couple of reasons. One reason was to control the amount of variance attributable to the secondary dimensions; within Level 2, the items as a set discriminated more strongly on the primary dimension than on the secondary dimensions, attributing more variance to the primary dimension than to the secondary dimensions. Another reason for the rescaling was so the overall

cluster effect was weak but the effect varied across the items such that some items had a more noticeable effect than other items; the item-level ICCs ranged from .01 to .32. This cluster-effect scenario was motivated by the PISA data that were analyzed for illustrative purposes in the main article. The rescaling of the item discriminations did not artificially restrict their range, as the final values ranged from 0.20 to 1.81 (see Table B1 for the rescaled values).

Regarding the overall intercepts (β_k) and the set of relative category intercepts ($\boldsymbol{\tau}_k$), first, the values for the direct category intercepts ($t_{kc} = \beta_k + \tau_{kc}$) were drawn. Then, they were transformed to β_k and $\boldsymbol{\tau}_k$. This approach ensured that the relative category intercepts monotonically advanced in a sufficient manner such that all of the items' categories were represented in the generated data (i.e., null categories did not occur for any of the items). The first (t_{k1}), second (t_2), and third (t_3) direct thresholds were randomly drawn from $U(-2.75, 0.50)$, $U(t_1 + 0.75, t_1 + 2.25)$, and $U(t_2 + 0.75, t_2 + 2.25)$, respectively. The generating value for item k's overall intercept was obtained by taking the mean of the direct intercepts for that item, $\beta_k = (t_{k1} + t_{k2} + t_{k3})/3$. The relative intercept for category c was obtained by taking the difference between the overall intercept and the direct intercept for that category, $\tau_{kc} = \beta_k - t_{kc}$. Table B1 also includes the values used for the overall and relative category intercepts to generate the data.

Table B1

The Values Used for the Item Discriminations, Overall Item Intercepts, and Relative Category Intercepts to Generate the Data for the Simulations

		Level 3	Level 2					Overall Intercept	Relative Category Intercepts		
Item	ICC_{k1}	$\tilde{\alpha}_{k1}$	α_{k1}	α_{k2}	α_{k3}	α_{k4}	α_{k5}	β_k	τ_{k1}	τ_{k2}	τ_{k3}
1	0.03	0.24	1.43	1.12				0.44	−0.98	−0.10	1.08
2	0.32	0.53	0.78	0.85				−0.19	−0.90	−0.08	0.97
3	0.09	0.34	1.07	1.02				0.67	−1.04	−0.15	1.19
4	0.02	0.25	1.63	0.96				0.50	−1.08	−0.05	1.13
5	0.16	0.34	0.79	1.31				−1.13	−1.09	0.12	0.98
6	0.02	0.21	1.64	1.30				1.16	−1.01	−0.15	1.15
7	0.11	0.37	1.03	0.63				0.15	−1.14	−0.12	1.26
8	0.02	0.21	1.61		0.52			−0.90	−1.28	−0.03	1.31
9	0.02	0.23	1.73		1.10			0.66	−1.40	0.09	1.31
10	0.02	0.21	1.75		1.28			0.05	−1.15	0.19	0.96
11	0.07	0.26	0.97		1.21			0.61	−0.95	0.09	0.86
12	0.03	0.21	1.30		0.65			0.71	−0.99	0.06	0.93
13	0.04	0.25	1.21		1.17			−0.07	−0.94	0.07	0.87
14	0.05	0.23	1.02		1.47			−1.39	−1.22	0.07	1.15
15	0.03	0.27	1.60			1.19		−0.27	−1.09	−0.15	1.24
16	0.09	0.36	1.15			0.89		0.43	−0.99	−0.11	1.10
17	0.01	0.21	1.81			1.03		0.11	−1.34	0.07	1.27
18	0.07	0.24	0.90			0.95		0.62	−0.92	−0.08	1.00
19	0.08	0.35	1.20			0.68		−1.60	−0.89	−0.06	0.95
20	0.10	0.37	1.12			1.08		1.30	−0.97	−0.21	1.18
21	0.12	0.37	1.00			1.32		−1.09	−0.90	0.08	0.83
22	0.10	0.28	0.85				1.00	−1.55	−1.06	0.10	0.97
23	0.24	0.36	0.64				1.60	0.87	−1.15	−0.14	1.29
24	0.02	0.20	1.57				0.91	1.50	−1.36	0.06	1.31
25	0.02	0.22	1.68				1.32	1.22	−1.00	0.02	0.98
26	0.08	0.27	0.93				0.61	0.46	−0.95	0.01	0.95
27	0.04	0.27	1.35				1.49	−1.65	−0.89	−0.13	1.02
28	0.05	0.28	1.16				0.58	0.58	−1.09	0.01	1.08

Note. An empty space indicates a value of 0. At Level 2 (i.e., the person level), Dimension 1 represents the primary dimension, and Dimensions 2 through 5 represent the secondary dimensions. At Level 3 (i.e., the cluster level), Dimension 1 represents the cluster dimension that corresponds to the Level 2 primary dimension.

Online Appendix C

This appendix includes details of simulations conducted to investigate the possibility of two other priors for the Bayesian correlated multilevel bifactor IRT model. The performances of these two versions of the model were compared to the performance of the model with informative priors (i.e., Model 1 in the main article, and this model remains Model 1 in this appendix).

In the first comparison model, which was Model 10 because the last model in the main article was Model 9, more conventional priors were assigned to the freely estimated item discriminations at Level 2 and Level 3. These conventional priors were lognormal distributions with means of 0 and standard deviations (*SD*s) of 1:

$$\alpha_{kd} \sim \text{logn}(0, 1), \text{ and}$$
$$\tilde{\alpha}_{kd} \sim \text{logn}(0, 1).$$

The prior distributions assigned to the remaining parameters were the same as those used in Model 1. Although these lognormal distributions were still informative priors, they were less so than the priors used in Model 1 because the *SD*s of the lognormal distributions were set to 1 in Model 1, whereas the *SD*s were set to 0.50 in Model 10. Therefore, this comparison model is referred to as the model with less informative priors.

In the second comparison model (Model 11), noninformative priors of uniform distributions ranging from a to b were assigned to all estimated parameters of the model, except the Level 2 and Level 3 latent trait dimensional positions. The Level 2 item discriminations were assumed to be distributed as

$$\alpha_{kd} \sim \begin{cases} U(0, 25) \text{ when item } k \text{ was the first discriminating item on dimension } d, \\ U(-25, 25) \text{ for all other estimated discriminations,} \end{cases}$$

and the Level 3 item discriminations were assumed to be distributed as

$$\tilde{\alpha}_{k1} \sim \begin{cases} U(0, 25) \text{ when item } k \text{ was the first discriminating item on the dimension.} \\ U(-25, 25) \text{ for all other discriminations.} \end{cases}$$

The lower bound of 0 for the first discriminating item on a dimension set the orientation of the metric (e.g., higher dimensional positions representing more of the measured trait); without this lower bound, the orientation of the metric could switch. The remaining parameters other than the latent trait dimensional positions were assigned the following priors:

$$\rho_{dd'} \sim U(-1, 1), \text{ where } d \neq d' \text{ and } d' > d > 1,$$
$$\beta_k \sim U(-25, 25) \text{ for all } k,$$
$$\tau_{kc} \sim U(-25, 25) \text{ for all } k \text{ and } c < m_k.$$

The latent trait dimensional positions were assigned the same priors as those used in Model 1.

Although these uniform distributions are referred to as noninformative priors, they are not truly noninformative. By setting the lower bounds of the distributions to 0 for the first discriminating items on their respective dimensions, the priors contribute information. The number of effective parameters, thus, is not equal to the number of estimated parameters (Gelman, Hwang, & Vehtari, 2014). These priors, however, are much less informative than those used in the original version of the model (Model 1) and the first comparison model (Model 10). Therefore, this model is referred to as the model with noninformative priors.

Additional Analyses

These comparison models (Models 10 and 11) were used to analyze the data generated for Correlational Condition 1 (the orthogonal condition) and Correlational Condition 4 (the orthogonal assumption was most violated) of the simulation study

reported in the main article. Additionally, these comparison models were used to analyze the PISA data related to interest in science. The results from these models were compared to those from the model with informative priors (Model 1). The total number of MCMC sampling iterations, the number of saved samples, and the thin rate were the same as those used in the simulation study reported in the main article.

Analysis of the Simulated Data

The LPMLs and BFs used to determine which model was favored most are in Table C1. In all conditions, the model with informative priors (Model 1) was at least strongly favored over the other models (based on the same criterion used in the main article; a BF greater than 6 indicated strong support). As for the recovery of the item discriminations, the bias plots in Figure C1 show the extent of the bias in the item discrimination estimates under Correlational Condition 4. Focusing on the model with noninformative priors (Model 11), when the sample size was 500, the Level 3 item discrimination estimates were strongly negatively biased (see Figure C1a). Moreover, the estimates for two items' discriminations on secondary dimensions showed sharp increases in positive bias relative to the other items (see Items 21 and 23 in Figure C1c; the bias was 0.78 for Item 23, so its marker does not appear in the figure). When the sample size was 1,000, the Level 3 item discrimination estimates were still negatively biased for this model, although the degree of bias for this sample size was less than that for the sample size of 500 (see Figure C1d).

Regarding the other two models, when the sample size was 500, most of the item discrimination estimates were similar for the model with informative priors (Model 1) and the model with less informative priors (Model 10). The exceptions were for a few of the Level 3 item discriminations, although the differences were small (e.g., for Items 1 through 9; see Figure C1a). When the sample size was 1,000, the item discrimination estimates were very similar for both models.

Analysis of the PISA Data

The LPMLs and BFs from the analysis of the PISA data are in Table C1 (see the "PISA" column). The model with informative priors (Model 1) was strongly favored over the other two models. Even though Model 1 was favored over the model with less informative priors (Model 10), the estimates from the two models were similar, a trend also observed in the simulations performed for this appendix. However, the trace plots depicting the Markov chains reveal the benefits of the informative priors used in Model 1; in general, the chains from Model 1 were more stable.

Figure C2 contains trace plots that show the differences in the stability of the chains. In this figure, the subplots in the first column (subplots a, c, e, and g) are trace plots related to Model 1, and the subplots in the second column (subplots b, d, f, and h) are trace plots related to Model 10. Each row (e.g., subplots a and b) are trace plots for the same parameter but from different models. The chains from Model 10 show greater fluctuation in the sampling process than those from Model 1. All chains, however, do not show differences regarding stability. The trace plots in Figure C3 (subplots a, b, c, and d) represent chains from both models that display similar levels of stability during the sampling process.

The greater fluctuation in the chains from Model 10 led to slightly larger variability in the posterior densities of the parameters associated with these chains. The posterior densities of the correlations (see Figure C3, subplots e, f, g, and h) provide examples of this issue. The densities for Model 10 (subplots f and h) have slightly larger variability (i.e., the densities are slightly flatter and wider) than those for Model 1 (subplots e and g). In terms of posterior summaries, the posterior mean and *SD* of the correlation between the secondary dimensions of "tasks" and "value" were respectively .374 and .048 for Model 1, and .386 and .056 for Model 10. For the correlation between the secondary dimensions of "value" and "activities," the posterior mean and *SD* were respectively .132 and .068 for Model 1, and .146 and .083 for Model 10. For both correlations, the posterior *SD*s from

Model 10 were slightly greater than those from Model 1.

Discussion

The recovery results of the additional simulations showed that using noninformative priors (Model 11) could lead to biased item discrimination estimates, and the bias for a couple of items indicated possible estimation instability for the sample size of 500. The instability could be because the data did not provide enough information to estimate the parameters, as the amount of information the data provides decreases as the sample size decreases. When the data provide an insufficient amount of information, adding information through the priors could help to obtain more accurate estimates, as Models 1 and 10 demonstrated for the sample size of 500; the estimates from these models did not show any sharp increases in bias as the estimates from Model 11 did.

In terms of the less informative priors, the slight fluctuation in the chains from the model might not be egregious enough to conclude that those priors are ineffective. Nevertheless, compared with Model 10, the model with informative priors (Model 1) resulted in more stable chains, estimates that were slightly less biased, and posterior densities with less variability. Given the results, favoring the informative priors over the less informative priors is reasonable. There are many other possible priors, however, and all of them cannot be explored in a single study. It is possible that other priors could lead to models that perform similarly to, or even outperform, Model 1.

References

Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, *24*(6), 997–1016.

Table C1

The Log-Predicted Marginal Likelihoods (LPMLs) With the Bayes Factors (BFs) on a 2 × Log Scale in Parentheses, From the Additional Simulations and the Analysis of the Program for International Student Assessment (PISA) Data Related to Interest in Science

	Simulations				Real Data
	Correlational Condition 1: Orthogonal		Correlational Condition 4: Greatest Violation		
	$N = 500$	$N = 1,000$	$N = 500$	$N = 1,000$	PISA
Three Levels: Unconstrained Discriminations					
Model 1: Informative Priors	−13,708	−27,256	−13,622	−27,163	−22,364
Model 10: Less Informative Priors	−13,717 (17.4)	−27,260 (8.0)	−13,631 (16.5)	−27,168 (10.5)	−22,372 (16.4)
Model 11: Noninformative Priors	−13,723 (30.1)	−27,266 (20.0)	−13,633 (22.1)	−27,173 (20.0)	−22,373 (18.2)

Note. For each model, the reported LPMLs and BFs within a sample size and correlational condition are the averages across the 25 generated data sets. The BFs indicate the extent to which the data supported Model 1 over the other models. Model 1 was the Bayesian correlated multilevel bifactor IRT model with informative priors, as presented in the main article. Model 10 was the model with less informative priors. Model 11 was the model with noninformative priors. All three models were multilevel bifactor IRT models, with the item discriminations unconstrained across levels and the secondary dimensions allowed to correlate with each other. Correlational Conditions 1 and 4 are the same as those in the simulation study reported in the main article.

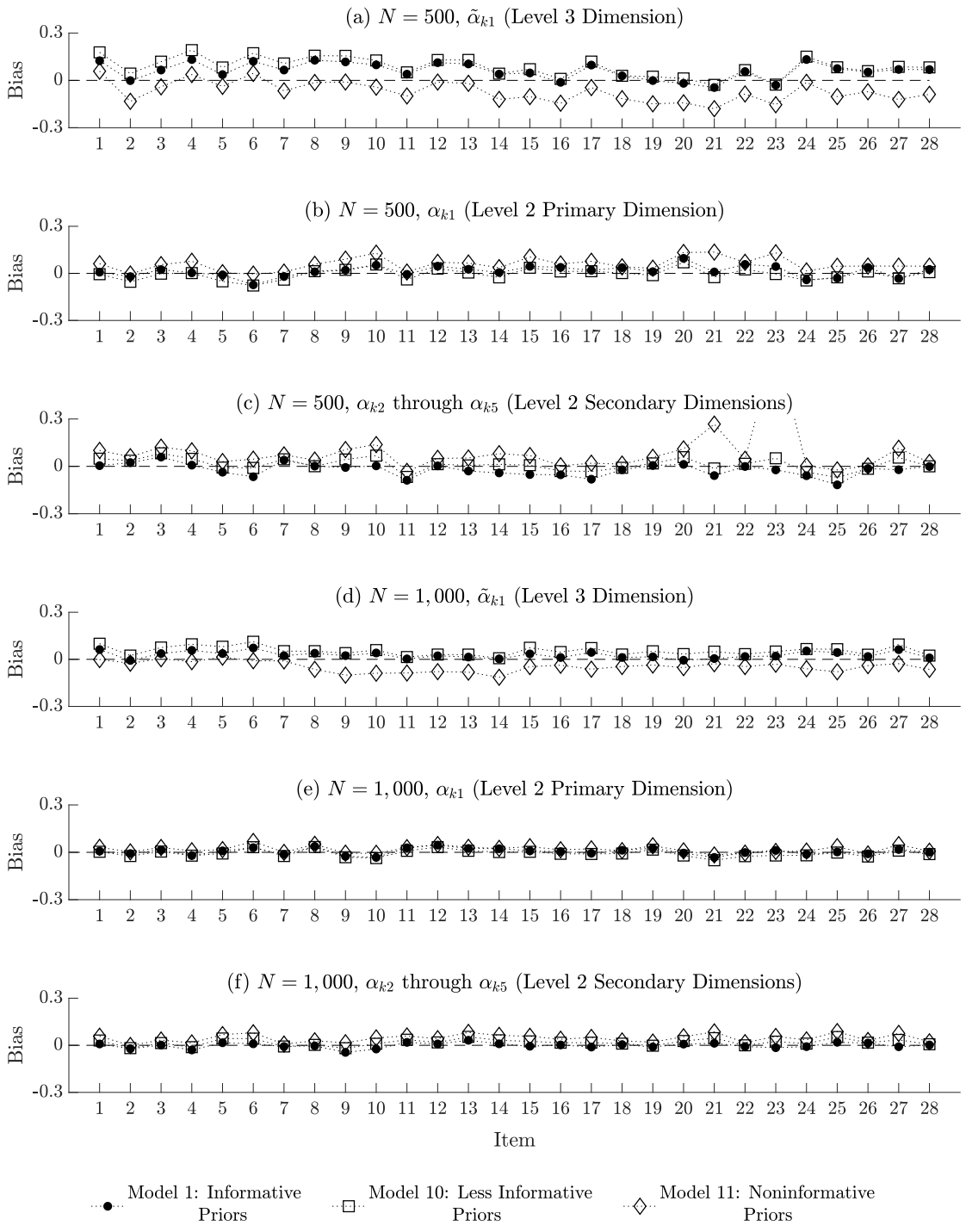


Figure C1. Bias plots detailing the recovery of the item discriminations in Correlational Condition 4 (the condition in which the orthogonal assumption was most violated). Subplots (a), (b), and (c) are for the sample size of 500; and subplots (d), (e), and (f) are for the sample size of 1,000. In each subplot, the items are represented along the horizontal axis, and the bias in the estimates is represented along the vertical axis.

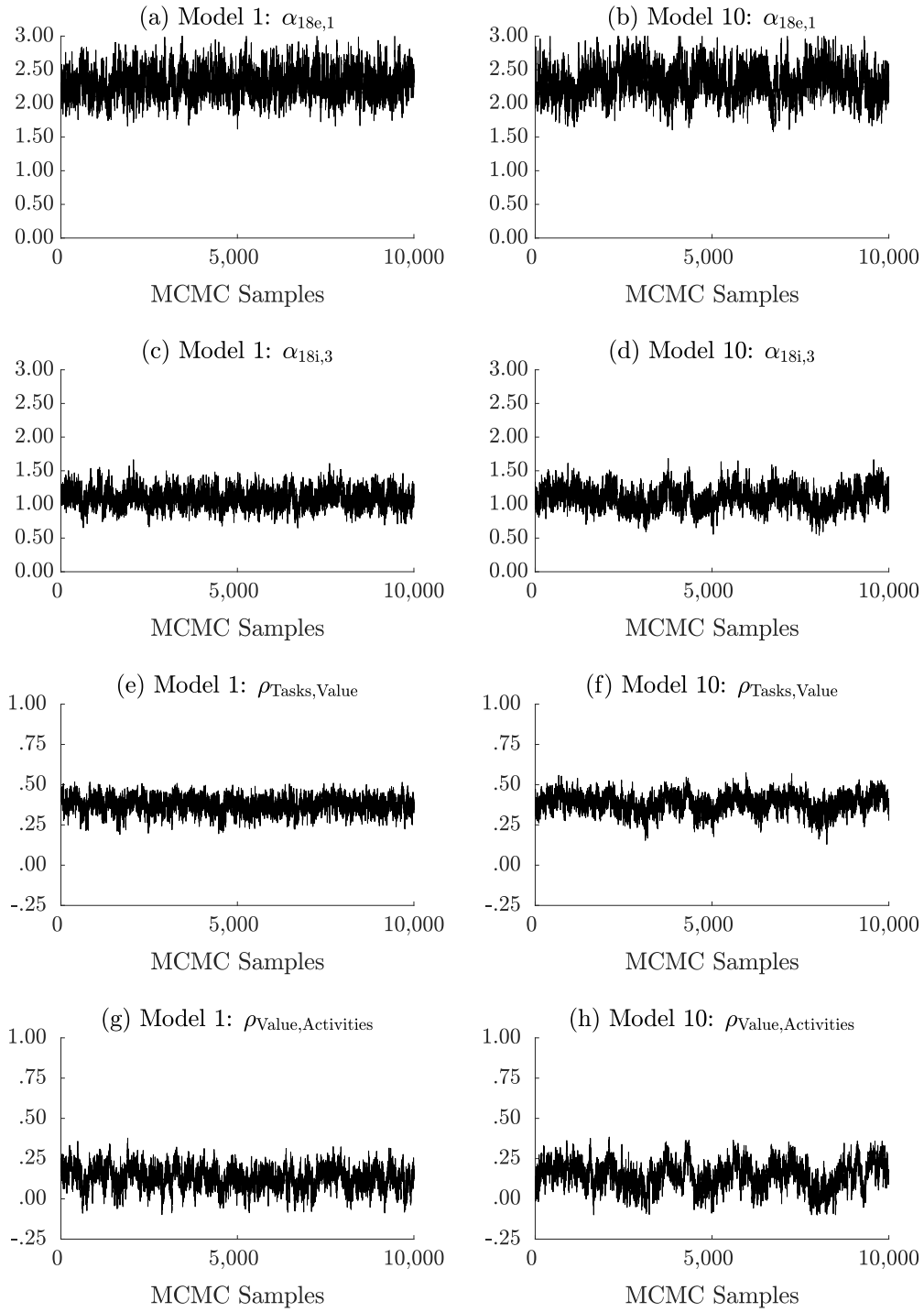


Figure C2. The following are trace plots from the models with informative priors (Model 1) and less informative priors (Model 10), produced during the analysis of the Program for International Student Assessment 2006 data related to interest in science. Subplots (a) and (b) are for an item's discrimination on the Level 2 primary dimension; subplots (c) and (d) are for an item's discrimination on a secondary dimension; subplots (e) and (f) are for the correlation between the secondary dimensions of tasks and value; and subplots (g) and (h) are for the correlation between the secondary dimensions of value and activities.

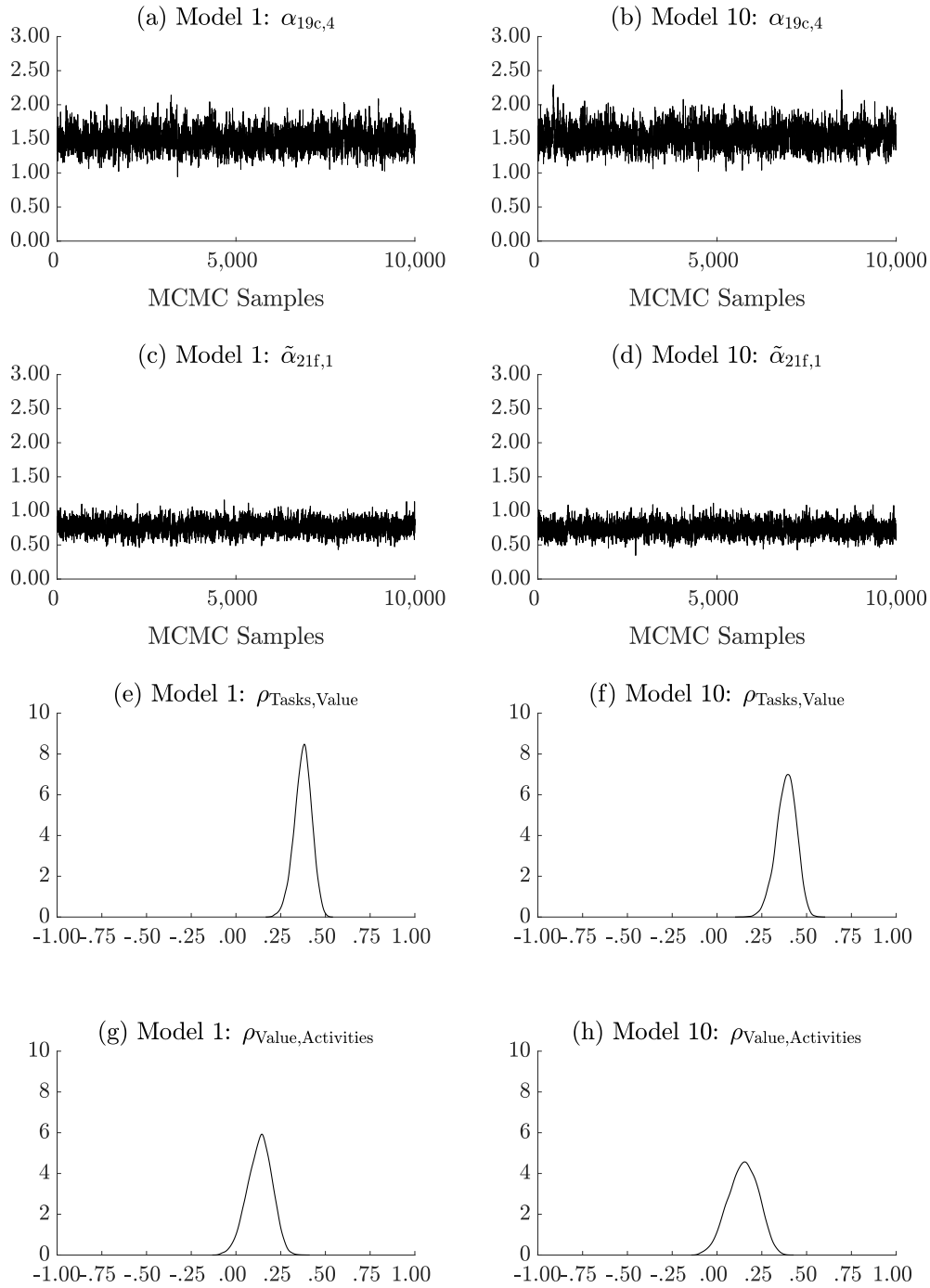


Figure C3. The following are trace plots and posterior densities from the models with informative priors (Model 1) and less informative priors (Model 10), produced during the analysis of the Program for International Student Assessment 2006 data related to interest in science. Subplots (a) and (b) are trace plots for an item's discrimination on a Level 2 secondary dimension; subplots (c) and (d) are trace plots for an item's discrimination on the Level 3 dimension; subplots (e) and (f) are posterior densities for the correlation between the secondary dimensions of tasks and value; and subplots (g) and (h) are posterior densities for the correlation between the secondary dimensions of value and activities.

Online Appendix D

This appendix includes details of simulations examining the performances of multilevel confirmatory factor analytic (MCFA) models, using the mean- and variance-adjusted weighted least squares estimation method (WLSMV; B. O. Muthén & Asparouhov, 2013). As it was noted in the primary article, an MCFA model based on a correlated multilevel bifactor structure is similar to a version of the Bayesian correlated multilevel bifactor IRT model based on the graded response model (GRM; Samejima, 1969), specified with a probit link function, rather than how the model was presented in the main article, which was based on the generalized partial credit model (GPCM; Muraki, 1992) and a logit link function. For the simulations reported in this appendix, all data were generated in a manner appropriate for an MCFA model.

These simulations used the sample sizes and two of the correlational conditions of the simulation study reported in the main article. The sample sizes were 500 and 1,000 (with cluster sizes of 65 and 130, respectively), and the correlational conditions were those in which the orthogonal assumption was met and most violated (Correlational Conditions 1 and 4, respectively). The data were generated using a multilevel bifactor IRT model based on the GRM, specified with a probit link function. The values in Online Appendix A were rescaled to the probit metric, using a factor of 1.7, and then used to generate the data. Although relative category intercepts have different meanings under the GPCM and the GRM, the values for the category intercepts were drawn to be ordered for the GPCM version of the model and were therefore still applicable for the GRM version. One hundred data sets were generated for each condition.

Two MCFA models were used to analyze the data. MCFA 1 and MCFA 2 were based on a multilevel bifactor structure, with both allowing the cluster effect to vary across the items by estimating a separate set of factor loadings at the between-cluster level (i.e., the cluster group level) and the within-cluster level (i.e., the person level). The difference between these models was that the former allowed the secondary factors to correlate with

each other and the latter restricted the secondary factors to be orthogonal to each other. All generated data sets were analyzed using the two models, and all analyses were performed using MPLUS 8.1 (L. K. Muthén & Muthén, 1998–2017).

When WLSMV is used, the items at the cluster level are treated as random intercepts, and so $K(K-1)/2$ sets of bivariate integrations take place (with K denoting the number of items). A quadrature based algorithm is used to perform these integrations (B. O. Muthén & Asparouhov, 2013). Therefore, a number of quadrature points must be selected. Seven, 15, and 25 quadrature points were used in these simulations. Seven points was included because it is the default value in MPLUS, and the larger numbers were selected to investigate whether more quadrature points improved estimation convergence.

Estimation convergence was determined in two ways. One way was the number of times the estimation process terminated without any errors. The second way was the number of times the estimation process terminated without any errors and did not produce any standard errors (SEs) for the factor loadings that were greater than 1.0. This threshold of 1.0 is arbitrary, but because the SEs for the factor loadings were generally below 0.30, an SE greater than 1.0 is a sharp increase in size relative to other SEs. Flagging these sharp increases is relevant because such behavior could signal estimation instability. Moreover, SEs are used in statistical significance testing of factor loading estimates to determine whether the estimates are different from 0. Thus, ensuring whether the SEs are trustworthy is important.

Table D1 includes the summary of the number of times the estimation process converged for a model. Of primary interest was the percentage of times the estimation process terminated without errors and the SEs for the factor loadings were less than 1.0 for reasons previously noted. For Correlational Condition 1 and the sample size of 500, the convergence rates for MCFA 1 and MCFA 2 were only 63% and 62%, respectively, with seven quadrature points. The convergence rates improved to 92% for both models when the number of quadrature points increased to 15. Increasing the number of quadrature points

to 25, however, did not lead to any further improvements in convergence rates. Similar trends were observed when the sample size was 1,000. That is, both models displayed similar convergence rates for each number of quadrature points. Additionally, the most noticeable improvement in convergence rates occurred when the number of quadrature points increased from seven to 15; almost perfect convergence rates were observed with 15 quadrature points (99% for both models). Increasing the number of quadrature points from 15 to 25 led to convergence rates of 100%, although this was only a 1% increase.

In Correlational Condition 4 (the greatest violation in the orthogonal assumption), regardless of the sample size, the most noticeable improvement in convergence rates was when the number of quadrature points increased from seven to 15. The convergence rates, however, did not improve when the number of quadrature points increased from 15 to 25. In fact, the highest convergence rates were observed with 15 quadrature points. When the sample size was 500, the convergence rates for MCFA 1 and MCFA 2 were 88% and 85%, respectively, and when the sample size was 1,000, the convergence rates for both models were 94%. These convergence rates were below 100%, even when the sample size was 1,000. The reason for the lack of perfect convergence rates could be that WLSMV does not incorporate prior information (other than that for the latent distribution) to estimate the models' parameters. As suggested in Online Appendix C, a lack of prior information could lead to estimation issues with smaller samples when MCMC is used, and the lack of prior information could be causing similar issues for WLSMV.

These findings should be interpreted with caution, however. These simulations did not examine the causes of these estimation difficulties because exploring them was beyond the aims of the main article. Sample sizes larger than 1,000 might be needed for WLSMV to converge at a higher rate when working within a dual-dependent context. Such sample sizes were not investigated in this appendix because the sample size for the PISA data analyzed in the main article was between 500 and 1,000. Understanding when WLSMV is appropriate for dual-dependent contexts could be beneficial because it could be less

computationally burdensome than Markov chain Monte Carlo methods.

References

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*(2), 159–176.

Muthén, B. O., & Asparouhov, T. (2013). Item response modeling in Mplus: A multi-dimensional, multi-level, and multi-timepoint example. In W. J. van der Linden & K. Hambleton (Eds.), *Handbook of item response theory: Models, statistical tools, and applications*. Boca Raton, FL: Chapman and Hall/CRC Press.

Muthén, L. K., & Muthén, B. O. (1998–2017). *Mplus user's guide*. Eighth Edition. Los Angeles, CA: Muthén and Muthén.

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

Table D1

The Convergence Rates (in Percentages) for the Multilevel Confirmatory Factor Analytic (MCFA) Models With Unconstrained Factor Loadings Across Levels

			Quadrature Points = 7		Quadrature Points = 15		Quadrature Points = 25	
Condition	*N*	Model	No Errors	No Errors and Large SEs	No Errors	No Errors and Large SEs	No Errors	No Errors and Large SEs
1: Orthogonal	500	MCFA 1	93	63	92	92	93	92
		MCFA 2	93	62	92	92	93	92
	1,000	MCFA 1	97	86	99	99	100	100
		MCFA 2	97	86	99	99	100	100
4: Greatest Violation	500	MCFA 1	94	71	89	88	89	88
		MCFA 2	95	66	88	85	86	83
	1,000	MCFA 1	94	81	94	94	94	94
		MCFA 2	94	77	94	94	93	93

Note. SEs = standard errors. "No errors" indicates the percentage of times the estimation process terminated without any errors. "No errors and large SEs" indicates the percentage of times the estimation process terminated without any errors and did not produce any SEs for the factor loadings that were greater than 1.0. These correlational conditions matched those of the simulation study reported in the main article. MCFA 1 and MCFA 2 allowed the cluster effect to vary across the items. The difference between the two models was that the former allowed the secondary factors to correlate with each other and the latter model treated the secondary factors as orthogonal to each other. The convergence rates were out of 100 simulated data sets.