Does Model Misspecification Lead to Spurious Latent Classes? An Evaluation of Model Comparison Indices

Ying-Fang Chen and Hong Jiao

1 Introduction

In recent decades, researchers have paid increasing attention to developing extended item response theory (IRT) models. These models were developed primarily because of the need to resolve the violation of the strong assumptions of IRT models, to more clearly reflect the nature of real-world testing scenarios, and to generate more accurate estimates of model parameters. One such extension is mixture IRT modeling (Kelderman and Macready 1990; Mislevy and Verhelst 1990; Rost 1990), which integrates an IRT or a Rasch model with latent class analysis (LCA) (Dayton 1999; McCutcheon 1987) to accommodate heterogeneity in examinee population.

In educational or psychological assessments, examinees/respondents may not be qualitatively homogeneous in terms of item response patterns. If examinees form a mixture of latent subgroups but a single latent population is assumed, the assumption of local independence in IRT models is violated. When such violations are not taken into account, the estimation of model parameters can be affected. For these reasons, the mixture modeling approach—which can identify the number of latent classes as well as describe multiple latent classes in the examinee population—has been progressively used in assessments (e.g., Cohen and Bolt 2005; De Ayala et al. 2002; Finch and Pierson 2011; Maij-de Meij et al. 2010; Mislevy and Verhelst 1990; Samuelsen 2005; Smith et al. 2012).

In the framework of mixture IRT modeling, the most frequently used mixture model is the mixture Rasch model (MRM) (Rost 1990). The MRM combines the Rasch measurement model (Rasch 1960) and LCA, allowing for multiple latent populations. For example, two examinees who have identical ability levels but belong

Y.-F. Chen (✉) • H. Jiao
Department of Human Development and Quantitative Methodology,
University of Maryland, College Park, MD 20742, USA
e-mail: pie@umd.edu; hjiao@umd.edu

R.E. Millsap et al. (eds.), *New Developments in Quantitative Psychology*,
Springer Proceedings in Mathematics & Statistics 66, DOI 10.1007/978-1-4614-9348-8_22,

to qualitatively heterogeneous subgroups are allowed to perform differentially on items (i.e., different item difficulties). The unconditional probability of a response vector can be expressed as

$$\mathrm{P}\left(\mathrm{x}\middle|\theta\right) = \sum_{\mathrm{c}=1}^{\mathrm{C}} \pi_{\mathrm{c}} \mathrm{P}\left(\mathrm{x}\middle|\theta, \mathrm{c}\right), \tag{1}$$

and the conditional probability of success given the latent class membership and model parameters for a specific latent class in the MRM is

$$\mathrm{P}\left(\mathrm{x} = 1\middle|\theta_{\mathrm{c}}, \mathrm{c}\right) = \frac{\exp\left(\theta_{\mathrm{jc}} - \beta_{\mathrm{ic}}\right)}{1 + \exp\left(\theta_{\mathrm{jc}} - \beta_{\mathrm{ic}}\right)}, \tag{2}$$

where $x = (x_1, \ldots, x_{\mathrm{I}})$ represents the response vector, π_{c} is the mixing proportion with a constraint $\sum \pi_c = 1$, β_{ic} is the difficulty for item i conditional on latent class c $(c = 1, \ldots, \mathrm{C})$, and θ_{jc} denotes the ability parameter for an examinee j in latent class c. For each latent class, the Rasch model is assumed. Both item and ability parameters are conditional on a discrete latent class. For scale identification, item difficulties within a class are usually constrained with $\sum_i^I \beta_{ic} = 0$ (Rost 1990). If a one-class solution is suggested for the data, the MRM is reduced into the Rasch model.

Given that the estimation of model parameters in mixture modeling depends on the identification of latent class membership, accurately extracting latent classes is important. If errors in latent class extraction occur, the estimation and interpretation of model parameters will be accordingly biased (e.g., Alexeev et al. 2011; Cho et al. 2012; Li et al. 2009). Alexeev et al. (2011) have recently demonstrated that model misspecification contributed to the creation of spurious latent classes in the MRM. The authors applied the MRM to the data generated as 2PLM. Their simulation results showed that at least two classes were extracted despite the fact that only one class was simulated. Their findings suggest that the extraction of latent classes did not result from examinee heterogeneity, but from model misspecification in the MRM.

Alexeev et al. (2011) provided valuable insights into potential nuisance sources that cause the formation of spurious latent classes. A major limitation in their simulation study, however, is that only varying item discrimination was considered as a single source of model misspecification (i.e., only 2PLM data were generated). Consequently, little is known as to whether their findings are generalizable to other testing conditions, such as model misspecification due to an addition of item guessing and/or slipping parameters that are not represented in the Rasch model.

In addition to item difficulty and discrimination, guessing and slipping parameters can also characterize items. A logistic IRT model that contains four item parameters (4PLM) (Barton and Lord 1981) is expressed as

$$P\left(x = 1\middle|\alpha_i, \beta_i, \gamma_i, \lambda_i, \theta_j\right) = \gamma_i + \frac{\lambda_i - \gamma_i}{1 + \exp\left(-\alpha_i\left(\theta_j - \beta_i\right)\right)}, \tag{3}$$

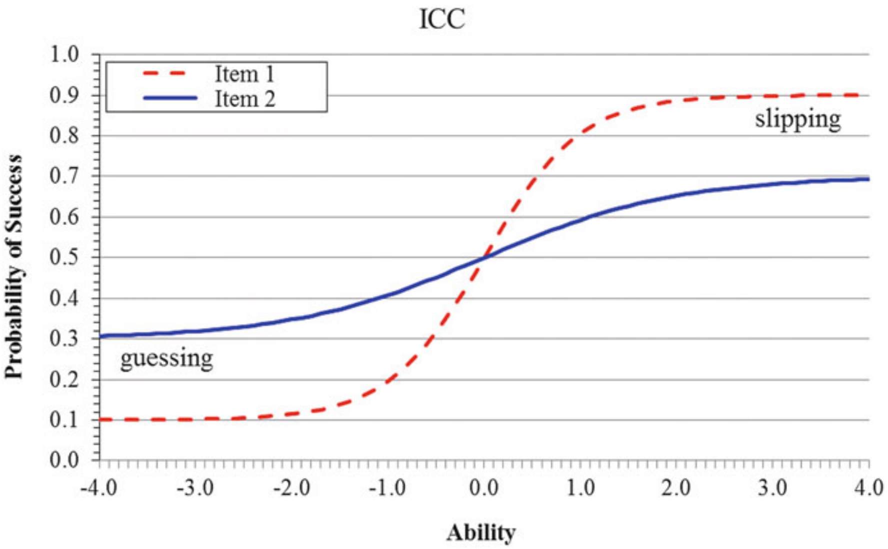

Fig. 1 Two 4PLM ICCs (item 1: $\alpha = 2.0$, $\gamma = 0.1$, $\lambda = 0.9$; item 2: $\alpha = 1.0$, $\gamma = 0.3$, $\lambda = 0.7$)

where α_i, β_i, γ_i, λ_i, and θ_i represent item discrimination, item difficulty, item guessing, item slipping, and ability parameter, respectively. In the plot of item characteristic curves (ICCs), item difficulty and discrimination represent the location and slope of an ICC, and the guessing parameter refers to the lower asymptote of an ICC and slipping is the upper asymptote of the ICC. Figure 1 illustrates two 4PLM ICCs, in which item 2 ($\alpha = 1.0$, $\gamma = 0.3$, $\lambda = 0.7$) exhibits stronger degrees of item guessing and slipping but weaker item discrimination than does item 1 ($\alpha = 2.0$, $\gamma = 0.1$, $\lambda = 0.9$). The probability of success in the 4PLM ranges from the lower asymptote to the upper asymptote across the ability continuum (Fig. 1). A guessing parameter is defined as the probability of a correct response by a very low-ability examinee. It can result from the use of a multiple choice format (e.g., the probability of success is 0.25 for a four-option item if a respondent guesses) and can be affected by flaws in an item (e.g., clues hidden in the item options or distractions are unattractive). A slipping item parameter is defined as the probability of an incorrect response to an item by a high-proficiency examinee. For some items, a high-ability examinee may unintentionally misread a question or overthink an easy item in a unique or creative manner. In a computer-based testing scenario, an item slipping effect may occur because of a special item response interface (Rulison and Loken 2009).

In the psychometric literature, the 3PLM with guessing is more prevalently used than IRT models with slipping. This predominance may be attributed to the fairly limited software available for estimating slipping parameters. Item slipping effects have been found to improve model-data fit and the accuracy of model parameter estimates in many empirical applications. For example, Barton and Lord

(1981) fitted the 4PLM and 3PLM to several large-scale assessment data sets and found that the former provided better model-data fit to the SAT verbal and math sections. In Loken and Rulison (2010), the 4PLM provided better fit (than did the 2PLM and 3PLM) to the delinquency data that were extracted from the large-scale *Monitoring the Future* (MTF) national survey (Johnston et al. 2006); in particular, the 4PLM yielded more information about moderate delinquency levels. Barton and Lord also demonstrated the adverse consequences of fitting the 3PLM and 2PLM to 4PLM data. Jiao et al. (2011) demonstrated how a 3PLM with slipping (i.e., 3PLM-λ) better fit a cognitive psychological test. Yen et al. (2012) indicated that in computerized adaptive testing scenarios, the 4PLM improved the ability estimates for a national sample data set that was drawn from the *English Ability Test* for college entrance in Taiwan. The above-mentioned studies suggest the practical need for and importance of including slipping effects to improve overall model-data fit and accuracy of parameter estimates.

The (mixture) Rasch model assumes no guessing or slipping effects. Therefore, model misspecification can also occur if one fits the (mixture) Rasch model to data with item guessing and/or slipping. To more comprehensively investigate the psychometric issue of the over-extraction of latent classes arising from model misspecification, the current study aims to examine whether the violation of assumptions regarding item discrimination, as well as guessing and slipping parameters, in the applications of the Rasch model causes the artificial extraction of latent classes. This research is expected to provide a more thorough discussion of the concerns about the over-extraction of latent classes. To sum up, this study intends to answer the following questions:

1. Which of the model-fit indices better selects the correct number of latent classes?
2. Does model misspecification cause the extraction of spurious latent classes in the MRM?
3. Do sample size and test length contribute to the extraction of spurious latent classes in the MRM?
4. How latent classes are extracted in real data applications?

2 Method

A simulation study was conducted to examine whether the extraction of spurious latent classes can be attributed to model misspecification. This research also explored the extraction of latent classes under real data scenarios. The succeeding section introduces the simulation design, real data sources, and data analysis methods used in this work.

Table 1 Simulation design

Manipulated factors	Levels
Item discrimination	1.0, 2.0
Item guessing	0.1, 0.3
Item slipping	0.7, 0.9
Test length	20, 40
Sample size	500, 1,000, 3,000

2.1 Simulation Study

This study was designed as a $3 \times 2 \times 2 \times 2 \times 2$ experimental design, in which sample size, test length, and magnitude of item discrimination, guessing, and slipping were manipulated. For each condition, 100 replications were simulated. Data were generated under a unidimensional IRT 4PLM [see Eq. (3)]. The data matrix comprises 500, 1,000, or 3,000 examinees' responses to 20 or 40 dichotomously scored items. Ability and item difficulty parameters were simulated from a standard normal distribution with a mean of 0 and a variance of 1.

Model misspecification in the Rasch model was manipulated using varying item discrimination and incorporating item guessing parameters (i.e., γ is greater than 0) and slipping parameters (i.e., λ is smaller than 1). Two levels of item discrimination, namely, 1.0 and 2.0, were used to represent low and high discrimination, respectively; these levels are identical to those adopted in previous studies (i.e., Emons et al. 2004; Li et al. 2009). Two levels of guessing (i.e., 0.1 and 0.3) and slipping effects (i.e., 0.7 and 0.9) were also manipulated as model misspecification factors. The extent of slipping was manipulated on the basis of slipping parameter estimates observed in previous empirical studies (i.e., $\lambda = 0.72$–0.89, Loken and Rulison 2010; $\lambda = 0.565$–0.998, Jiao et al. 2011). A discrimination of 2.0, a guessing level of 0.3, or a slipping level of 0.7 represents a strong violation of the Rasch model; that is, in the Rasch model, item discrimination is equal to 1, guessing is equal to 0, and slipping is equal to 1. The specifications used in the simulation study are summarized in Table 1.

Item response data were then analyzed with the MRM, which incorporates only difficulty parameters in the model [see Eqs. (1) and (2)]. The software *mdltm* developed by von Davier (2005) was used for estimation; it applies marginal maximum likelihood estimates with an expectation-maximization algorithm. Given that more than two extracted classes (e.g., two classes, three classes, and so on) indicate the presence of spurious latent classes (i.e., in the data generation, one latent class was simulated), this study reports the percentages of replications that suggest multiple-class solutions in the data. The outcome statistics for evaluating model-data fit are Akaike information criterion (AIC; Akaike 1974), Bayesian information criterion (BIC; Schwarz 1978), corrected Akaike information criterion (AICc; Burnham and Anderson 2002), and sample-size adjusted Bayesian information criterion (SABIC; Sclove 1987). These statistics are expressed as Eqs. (4)–(7). The

AIC and BIC measures were output from *mdltm*, and SABIC and AICc—which have penalties greater for a large number of parameters or small samples—were computed from Eqs. (6) and (7). The outcome statistics were computed across 100 replications:

$$\mathrm{AIC} = -2\mathrm{LL} + 2\mathrm{k}, \tag{4}$$

$$\mathrm{BIC} = -2\mathrm{LL} + \mathrm{k}\ln(\mathrm{N}), \tag{5}$$

$$\mathrm{AICc} = \mathrm{AIC} + \frac{2\mathrm{k}(\mathrm{k}+1)}{\mathrm{N}-\mathrm{k}-1}, \tag{6}$$

$$\mathrm{SABIC} = -2\mathrm{LL} + \mathrm{k}\ln\left(\frac{\mathrm{N}+2}{24}\right), \tag{7}$$

where LL = log-likelihood, k = number of parameters, and N = sample size.

2.2 *Empirical Examples*

Two real data sets were used. The first was extracted from a standardized large-scale assessment—the Progress in International Reading Literacy Study (PIRLS) 2006 (PIRLS 2006 Assessment 2007) conducted by the International Association for the Evaluation of Educational Achievement (Green et al. 2009; Mullis et al. 2007). PIRLS is designed to measure the reading comprehension abilities of fourth grade students. The extracted sample data set contains 1,398 examinees' responses to 21 items, in which the items were originally constructed under the 2PL or the 3PL model.

The second data set was extracted from the 2005 national MTF survey of 12th grade students (Johnston et al. 2006). The extracted data set contains 2,463 examinees' responses to 14 self-report questions of delinquency according to a 5-point Likert scale (i.e., students reported the frequency of delinquency acts; 1 = not at all to 5 = five or more times within the past year). The item responses were re-coded in binary format (i.e., 1 = at least once, 0 = never). Loken and Rulison (2010) demonstrated that the 4PLM satisfactorily fit this data set; the estimates of slipping parameters ranged from 0.72 to 0.89, which implicitly suggests that individuals at high levels of delinquency did not necessarily commit all delinquency acts. The data analysis procedure for the empirical examples is identical to that implemented in the above-mentioned simulation study.

3 Results

3.1 Simulation Study

Table 2 shows the average hit rates of latent class selection under each simulated condition, while Table 3 reveals the percentages of replications that extracted spurious latent classes in the MRM. The performance ranking of the model-fit indices followed the order BIC, SABIC, AICc, and AIC, with overage hit rates of 97.10 %, 86.30 %, 75.90 %, and 70.50 %, respectively (Table 2). Among the model-fit indices, BIC exhibited the best performance in selecting the correct number of latent classes, particularly under small- (i.e., sample size = 500) and moderate-sized (i.e., sample size = 1,000) samples (i.e., average hit rate = 100 %). BIC showed a relatively satisfactory and consistent performance across all simulated conditions (average hit rate = 91.25–100 %). By contrast, AIC produced the worst model selection, particularly when item discrimination was violated in the Rasch model (i.e., $\alpha = 2.0$) and when sample sizes increased (Table 3).

The inclusion of item guessing and slipping parameters did not contribute to the extraction of spurious latent classes in the MRM (Table 3). However, the model misspecification resulting from item discrimination was an influential factor for such extraction. At an item discrimination of 1 (i.e., as the constraint of item discrimination in the Rasch model), spurious latent classes were imperceptibly extracted even under item guessing and slipping effects.

Table 2 Average hit rates for latent class selection under simulated conditions (%)

	AIC	BIC	SABIC	AICc
Overall	70.50	97.10	86.30	75.90
Sample sizes				
500	94.63	100.00	100.00	100.00
1,000	65.25	100.00	96.19	74.19
3,000	51.63	91.25	62.56	53.63
Test lengths				
20	74.04	99.25	87.54	78.79
40	66.96	94.92	84.96	73.08
Item discrimination				
1.0	99.29	100.00	100.00	99.79
2.0	41.71	94.17	72.50	52.08
Item guessing				
0.1	64.83	94.71	80.83	71.46
0.3	76.17	99.46	91.67	80.42
Item slipping				
0.7	75.33	95.92	89.33	80.71
0.9	65.67	98.25	83.17	71.17

Table 3 Percentages of replications that extracted spurious classes in the mixture Rasch model (%)

Sample sizes				Fit indices											
				AIC			BIC			SABIC			AICc		
Item	α	γ	λ	500	1,000	3,000	500	1,000	3,000	500	1,000	3,000	500	1,000	3,000
20	1	0.1	0.7	0	0	0	0	0	0	0	0	0	0	0	0
20	1	0.1	0.9	0	0	0	0	0	0	0	0	0	0	0	0
20	1	0.3	0.7	0	0	0	0	0	0	0	0	0	0	0	0
20	1	0.3	0.9	0	0	0	0	0	0	0	0	0	0	0	0
40	1	0.1	0.7	0	0	16	0	0	0	0	0	0	0	0	5
40	1	0.1	0.9	0	0	0	0	0	0	0	0	0	0	0	0
40	1	0.3	0.7	0	0	0	0	0	0	0	0	0	0	0	0
40	1	0.3	0.9	0	0	1	0	0	0	0	0	0	0	0	0
20	2	0.1	0.7	0	**69**	**100**	0	0	1	0	0	**100**	0	22	**100**
20	2	0.1	0.9	3	**93**	**100**	0	0	14	0	0	**100**	0	**74**	**100**
20	2	0.3	0.7	0	0	**58**	0	0	0	0	0	0	0	0	44
20	2	0.3	0.9	6	**94**	**100**	0	0	3	0	0	**99**	0	**69**	**100**
40	2	0.1	0.7	49	**100**	**100**	0	0	**97**	0	**56**	**100**	0	**99**	**100**
40	2	0.1	0.9	15	**99**	**100**	0	0	15	0	4	**100**	0	**85**	**100**
40	2	0.3	0.7	0	1	**99**	0	0	0	0	0	0	0	0	**93**
40	2	0.3	0.9	13	**100**	**100**	0	0	10	0	1	**100**	0	**64**	**100**

When item discrimination varied from one as a kind of model misspecification in the Rasch model, the extraction of spurious latent classes increased; in particular, all the indices tended to extract spurious latent classes as sample sizes and test lengths increased. Overall, BIC was the least influenced by model misspecification resulting from item discrimination. For example, when $\alpha = 1$, all the model-fit indices produced perfect or nearly perfect hit rates (i.e., average hit rate = 99.29–100 %). When $\alpha = 2$, however, the average hit rate of BIC remained high (94.17 %), whereas those of the other indices visibly decreased (41.71 %, 52.08 %, and 72.50 % for AIC, AICc, and SABIC, respectively). Across all model misspecification conditions (i.e., $\alpha = 2.0$, $\gamma = 0.1$ or 0.3, $\lambda = 0.7$ or 0.9), BIC resulted in high hit rates that ranged from 94.17 to 99.46 % (Table 2).

3.2 Empirical Examples

Tables 4 and 5 show the results of the latent class selection for the PIRLS and the MTF data sets, respectively. In the empirical examples, the model-fit indices generated conflicting results, making the determination of the number of latent classes difficult. For both real data sets, BIC suggested a one-class solution, whereas SABIC recommended a two-class solution and AIC and AICc selected three-class solutions. Given that BIC effectively selected the correct number of latent classes, the one-class solution was adopted for both real data sets. AIC, SABIC, and AICc

Table 4 Latent class selection for the PIRLS

	AIC (rank)	BIC (rank)	AICc (rank)	SABIC (rank)
One-class solution	30,299 (4)	30,414 (1)	30,299 (4)	30,344 (2)
Two-class solution	30,225 (3)	30,461 (2)	30,228 (2)	30,318 (1)
Three-class solution	30,212 (1)	30,569 (4)	30,219 (1)	30,353 (3)
Four-class solution	30,216 (2)	30,694 (3)	30,229 (3)	30,404 (4)

Table 5 Latent class selection for the MTF

	AIC (rank)	BIC (rank)	AICc (rank)	SABIC (rank)
One-class solution	18,963 (3)	19,271 (1)	18,966 (4)	19,103 (3)
Two-class solution	18,666 (2)	19,288 (2)	18,676 (2)	18,948 (1)
Three-class solution	18,642 (1)	19,578 (3)	18,665 (1)	19,066 (2)
Four-class solution	18,666 (2)	19,915 (4)	18,707 (3)	19,232 (4)

tended to select a model that had many latent classes in both empirical examples, echoing the results of the simulation study.

4 Summary and Discussion

This research investigated whether model misspecification results in an extraction of spurious latent classes in the MRM and assessed the effectiveness of model-fit indices in latent class selection. BIC was the most promising model-fit index for selecting the correct number of latent classes, whereas AIC and AICc tended to select a model with spurious latent classes in the MRM. Our findings are consistent with those of previous studies, in which BIC performed effectively and AIC functioned poorly in latent class selection (i.e., Alexeev et al. 2011; Cho and Cohen 2010; Cho et al. 2012, Li et al. 2009; Preinerstorfer and Forman 2012). As stated earlier, model parameter estimation considerably depends on the identification of latent class membership; therefore, inaccurate estimates of the number of latent classes can cause severe biases in model parameter estimates. Given that no consensus regarding the best indicator of latent class numeration in mixture modeling has been reached (Nylund et al. 2007), researchers and practitioners should be particularly cautious in choosing model-fit measures. As indicated by the current and previous findings, BIC is favorable for latent class selection in data.

In Alexeev et al. (2011), BIC extracted spurious latent classes under large sample sizes, long test lengths, and a distribution of item discrimination parameters that corresponds to a violation of uniform item discrimination in the MRM. In the current work, however, BIC reduces concerns over the extraction of spurious latent class resulting from model misspecification. More specifically, BIC extracted spurious latent classes only when the constraint of item discrimination in the MRM was violated in few large-sample conditions. Sample size and test length were not the primary influential factors in the current study. AIC, SABIC, and AICc over-extracted latent classes under large item discrimination combined with large sample sizes—a finding that corresponds with that of Alexeev et al. (2011). The slight differences in findings between the current study and Alexeev et al. (2011) may be due to the different estimation programs used (Mplus was used in the latter). An interesting and new finding from the present research is that the model misspecification resulting from item guessing and slipping effects did not contribute to the extraction of spurious latent classes in the MRM.

Finally, the empirical examples show that the model-fit indices presented inconsistent results, also an observed occurrence in previous studies that used real data (e.g., Cho and Cohen 2010; Li et al. 2009; Willse 2011). Inconsistent latent class selection in real data can be a serious concern because the true number of latent classes in real data is usually unknown. In our real data application, both sets of real data, which contain items best modeled with guessing and/or slipping parameters, did not lend themselves to additional latent classes (i.e., possibly spurious latent classes), as indicated by the BIC values.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723.

Alexeev, N., Templin, J., & Cohen, A. (2011). Spurious latent classes in the mixture Rasch model. *Journal of Educational Measurement*, *48*, 313–332.

Barton, M. A., & Lord, F. M. (1981). *An upper asymptote for the three-parameter logistic item response model* (Report No. RR-81-20). Princeton, NJ: Educational Testing Service.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). Berlin, Germany: Springer.

Cho, S.-J., & Cohen, A. S. (2010). A multilevel mixture IRT model with an application to DIF. *Journal of Educational and Behavioral Statistics*, *35*, 336–370.

Cho, Y., Jiao, H., & Macready, G. B. (2012). *Assessing the effects of different item parameter profiles in mixture Rasch models*. Paper presented at the meeting of the American Educational Research Association (AERA), Vancouver, Canada.

Cohen, A. S., & Bolt, D. M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement*, *42*, 133–148.

Dayton, C. M. (1999). *Latent class scaling analysis*. Thousand Oaks, CA: Sage.

De Ayala, R. J., Kim, S.-H., Stapleton, L. M., & Dayton, C. M. (2002). Differential item functioning: A mixture distribution conceptualization. *International Journal of Testing*, *2*, 243–276.

Emons, W. H. M., Sijtsma, K., & Meijer, R. R. (2004). Testing hypotheses about the person-response function in person-fit analysis. *Multivariate Behavioral Research*, *39*, 1–35.

Finch, H., & Pierson, E. (2011). A mixture IRT analysis of risky youth behavior. *Frontiers in Psychology, 2*, 98. doi:10.3389/fpsyg.2011.00098

Green, P. J., Herget, D., Rosen, J., & Provasnik, S. (2009). *User's guide for the Progress in International Reading Literacy Study* (*PIRLS*). Washington, DC: National Center for Education Statistics, Institute of Education Sciences.

Jiao, H., Macready, G., Zhu, J., & An, W. (2011). *A three parameter item response theory model with varying upper asymptote effects*. Paper presented at the meeting of the Psychometric Society, Hong Kong, China.

Johnston, L. D., Bachman, J. G., O'Malley, P. M., & Schulenberg, J. E. (2006). *Monitoring the future*: *A continuing study of American youth* (*12th-grade survey*)*, 2005* [Computer file]. ICPSR04536-v3. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2007-07-18. doi:10.3886/ICPSR04536.v3

Kelderman, H., & Macready, G. B. (1990). The use of loglinear models for assessing differential item functioning across manifest variables and latent examinee groups. *Journal of Educational Measurement*, *27*, 307–327.

Li, F., Cohen, A. S., Kim, S.-H., & Cho, S.-J. (2009). Model selection methods for mixture dichotomour IRT models. *Applied Psychological Measurement*, *33*, 353–373.

Loken, E., & Rulison, K. L. (2010). Estimation of a four-parameter item response theory model. *British Journal of Mathematical and Statistical Psychology*, *63*, 509–525.

Maij-de Meij, A. M., Kelderman, H., & van der Flier, H. (2010). Improvement in detection of differential item functioning using a mixture item response theory model. *Multivariate Behavioral Research*, *45*, 975–999.

McCutcheon, A. L. (1987). *Latent class analysis*. Beverly Hills, CA: Sage.

Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika*, *55*, 195–215.

Mullis, I. V. S., Martin, M. O., Kennedy, A. M., & Foy, P. (2007). *PIRLS 2006 International Report*. Boston, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Nylund, K. L., Asparouhov, T., & Muthen, B. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling*, *14*, 535–569.

PIRLS 2006 Assessment. (2007). *International Association for the Evaluation of Educational Achievement (IEA)*. Boston, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Preinerstorfer, D., & Forman, A. K. (2012). Parameter recovery and model selection in mixed Rasch model. *British Journal of Mathematical and Statistical Psychology*, *65*, 252–262.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Nielsen & Lydiche.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, *14*, 271–282.

Rulison, K. L., & Loken, E. (2009). I've fallen and I can't get up: Can high-ability students recover from early mistakes in CAT? *Applied Psychological Measurement*, *33*, 83–101.

Samuelsen, K. (2005). *Examining differential item from a latent class perspective* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3175148)

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.

Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, *52*, 333–43.

Smith, E. V., Jr., Ying, Y., & Brown, S. W. (2012). Using the mixed Rasch model to analyze data from the beliefs and attitudes about memory survey. *Journal of Applied Measurement, 13*, 23–40.

Von Davier, M. (2005). *Mdltm: Software for the general diagnostic model and for estimating mixture of multidimensional discrete latent traits models* [Computer software]. Princeton, NJ: Educational Testing Service.

Willse, J. T. (2011). Mixture Rasch models with joint maximum likelihood estimation. *Educational and Psychological Measurement*, *71*, 5–19.

Yen, Y.-C., Ho, R.-G., Laio, W.-W., Chen, L.-J., & Kuo, C.-C. (2012). An empirical evaluation of the slip correction in the four parameter logistic models with computerized adaptive testing. *Applied Psychological Measurement*, *36*, 75–87.