An Application of a Random Mixture Nominal Item Response Model for Investigating Instruction Effects

Hye-Jeong Choi, Allan S. Cohen, and Brian A. Bottge

Abstract The purpose of this study was to apply a random item mixture nominal item response model (RIM-MixNRM) for investigating instruction effects. The host study design was a pre-test-and-post-test, school-based cluster randomized trial. A RIM-MixNRM was used to identify students' error patterns in mathematics at the pre-test and the post-test. Instruction effects were investigated in terms of students' transitioning in error patterns. That is, we compared students' error patterns in the Enhanced Anchored Instruction (EAI) condition with students' error patterns in a business-as-usual (BAU) instructional condition following each instruction. We also compared error patterns of students with math disabilities and students without math disabilities following the two types of instruction.

Keywords Random item model • Mixture IRT model • Nominal responses model

1 Introduction

Mixture item response theory (MixIRT) models have been used for modeling population heterogeneity (e.g., Mislevy & Verhelst 1990; Rost 1990). Most mixture models consider only persons random but items fixed. Random item IRT models (De Boeck 2008), however, consider both item and person parameters as random. This is more appealing as (1) both items and persons are typically assumed to be random samples from some population and (2) treating both items and persons as random permits inclusion of covariates on both item and person parameters to help explain differences in both item and examinee parameters (Van den Noortgate, De Boeck, & Meulders, 2003; Wang 2011).

Also, to date, most MixIRT models have primarily focused on dichotomously or polytomously scored items. MixIRT models can be usefully applied to nominal

H.-J. Choi (✉) • A.S. Cohen
University of Georgia, Aderhold Hall 125, Athens, GA 30601, USA
e-mail: hjchoi1@uga.edu

B.A. Bottge
University of Kentucky, 222 Taylor Education Building, Lexington, KY, USA

L.A. van der Ark et al. (eds.), *Quantitative Psychology Research*, Springer
Proceedings in Mathematics & Statistics 167, DOI 10.1007/978-3-319-38759-8_4

responses items. For example, MixIRT models can effectively model to capture information about specific error patterns which individual distractors for multiple choice items may contain.

The purpose of this study was to apply a random item mixture nominal response model to an empirical data set for investigating instructional effects. An important benefit of such a model is that it is possible to explicitly model randomness of item and ability parameters as well as specific aspects of students' response patterns. We provide a brief description of the random item mixture nominal model (RIM-MixNRM) and a simulation study to evaluate the quality of the estimation method. Then, we provide an empirical example in which a RIM-MixNRM is applied to mathematic test data to investigate effects of an experimental instruction on students' error patterns on fractions computation.

2 A Random Item Mixture Nominal Response Model

The probability of selecting individual categories in an item with two or more nominal categories can be written as a linear function of item category and person parameters. Bock (1972), for instance, introduces a nominal model in which the probability of selecting category k of item i, $P_{ik}(\theta_j)$, is defined as a multinomial logistic function:

$$P_{ik}(\theta_j) = \frac{\exp(\lambda_{ik}\theta_j + \zeta_{ik})}{\sum_{k=1}^{K} \exp(\lambda_{ik}\theta_j + \zeta_{ik})}, \tag{1}$$

where

$i = 1, \ldots, n$ items,
$k = 1, \ldots, K$ response categories,
$j = 1, \ldots, N$ examinees,
ζ_{ik} denotes the intercept for category k of item i,
λ_{ik} denotes the slope for category k of item i, and
θ_j denotes the person parameter of person j.

Bolt, Cohen, and Wollack (2001) extended this model to a mixture nominal IRT model as a way of detecting heterogeneity in a population. In doing so, Bolt et al. (2001) included a class-specific category intercept parameter to specify the propensity of selecting a given category of item i for members of latent class g. The class-specific probability of a response is given by

$$P_{gik}(\theta_j) = \frac{\exp(\lambda_{ik}\theta_j + \zeta_{gik})}{\sum_{k=1}^{K} \exp(\lambda_{ik}\theta_j + \zeta_{gik})}, \tag{2}$$

with marginal probability

$$P_{ik}(\theta_j) = \sum_{g}^{G} \pi_g P_{gik}(\theta_j) \,, \tag{3}$$

where ζ_{gik} denotes the class-specific category intercept, $g = 1 \ldots, \mathrm{G}$ latent classes, and π_g mixing proportion ($\sum_g^G \pi_g = 1$). For resolving an indeterminacy for the item category parameters, constraints of $\sum_k^K \lambda_{ik} = 0$ and $\sum_k^K \zeta_{gik} = 0$ were set for all items and all classes.

Bolt et al. (2001) applied Markov chain Monte Carlo (MCMC) algorithm to estimate the model in a general hierarchical framework and a fully Bayesian approach as implemented in the computer software computer software WinBUGS. They used following conjugate priors:

$$
\begin{aligned}
& c_j \sim \mathit{Multinomial}(1|\pi_1, \ldots, \pi_G) \\
& \pi = (\pi_1, \ldots, \pi_G) \sim \mathit{Dirichlet}(\alpha_1, \ldots, \alpha_G) \\
& \theta_j|c_j = g \sim N(\mu_{\theta_g}, \sigma^2_{\theta_g}) \\
& \lambda_{ik} \sim N(\mu_\lambda, \sigma^2_\lambda) \\
& \zeta_{gik} \sim N(\mu_{\zeta_g}, \sigma^2_{\zeta_g}).
\end{aligned}
$$

In their model, however, item parameters were treated as fixed as in the conventional mixture item response models. In the current study, we extended their model to a model where both item and person parameters are treated as random.

3 Simulation Study

The simulation study described below was designed to examine the behavior of the RIM-MixNRM under practical testing conditions.

3.1 Simulation Design

Hundred sets of 20 four-choice item responses were simulated from a standard normal distributions, $N(0, 1)$. Six hundred examinees for three latent classes were simulated and mixing proportions were 0.33 and ability was generated as $N(0, 1)$ for each class. Item parameter estimates adopted from Bolt et al. (2001) were used to select item generating parameters. Generating values for model parameters are given in Table 1. As can be seen in Eq. (2), in this particular RIM-MixNRM, ζ_{gik} is the parameter to distinguish latent classes.

The parameters for hyperpriors were used: $\alpha_1 = \ldots = \alpha_G = 0.5$; $\mu_{\theta_g} \sim N(0, 1)$; $1/\sigma^2_{\theta_g} \sim \mathit{Gamma}(2, 4)$; $\mu_\lambda \sim N(0, 1)$; $1/\sigma^2_\lambda \sim \mathit{Gamma}(2, 4)$, $\mu_{\zeta_g} \sim N(0, 1)$; $1/\sigma^2_{\zeta_g} \sim \mathit{Gamma}(2, 4)$. These parameters were similar with ones used by Bolt et al. (2001) and only provided minimum information for each parameter. In addition to

Table 1 Item parameter for generating data sets for simulation study

	Slope				Threshold											
					Class 1				Class 2				Class 3			
	λ_1	λ_2	λ_3	λ_4	ζ_1	ζ_2	ζ_3	ζ_4	ζ_1	ζ_2	ζ_3	ζ_4	ζ_1	ζ_2	ζ_3	ζ_4
1	1.01	0.30	−0.14	−1.17	0.74	0.20	0.04	−0.98	0.74	2.70	0.04	−0.98	0.74	0.20	0.04	−0.98
2	0.89	−0.21	0.39	−1.07	0.44	0.30	−0.84	0.10	0.44	2.80	−0.84	0.10	0.44	0.30	−0.84	0.10
3	1.69	−0.29	−0.51	−0.89	−0.57	0.34	0.33	−0.10	−0.57	2.84	0.33	−0.10	−0.57	0.34	0.33	−0.10
4	0.99	−0.32	0.13	−0.80	−0.22	0.78	0.31	−0.87	−0.22	3.28	0.31	−0.87	−0.22	0.78	0.31	−0.87
5	0.73	−0.42	0.31	−0.62	−0.97	0.50	−0.93	1.40	−0.97	3.00	−0.93	1.40	−0.97	0.50	−0.93	1.40
6	1.62	−0.53	−0.54	−0.55	−1.86	0.02	2.51	−0.67	−1.86	2.52	2.51	−0.67	−1.86	0.02	2.51	−0.67
7	0.95	0.29	0.38	−1.62	−1.11	0.20	2.69	−1.78	−1.11	0.20	2.69	−1.78	−1.11	0.20	5.19	−1.78
8	1.06	−0.61	0.50	−0.95	0.42	−0.69	1.41	−1.14	0.42	−0.69	1.41	−1.14	0.42	−0.69	3.91	−1.14
9	1.20	−0.39	0.27	−1.08	1.25	−1.02	1.34	−1.57	1.25	−1.02	1.34	−1.57	1.25	−1.02	3.84	−1.57
10	0.91	0.12	0.78	−1.81	−0.77	1.36	−0.35	−0.24	−0.77	1.36	−0.35	−0.24	−0.77	1.36	2.15	−0.24
11	0.91	0.46	−0.43	−0.94	1.25	−0.16	−0.58	−0.51	1.25	−0.16	−0.58	−0.51	1.25	−0.16	1.92	−0.51
12	1.42	0.03	0.34	−1.79	0.16	0.35	0.22	−0.73	0.16	0.35	0.22	−0.73	0.16	0.35	2.72	−0.73
13	1.09	−0.23	−0.32	−0.54	1.63	−0.84	−0.12	−0.67	1.63	−0.84	−0.12	−0.67	1.63	−0.84	2.38	−0.67
14	1.19	−0.24	0.36	−1.31	−2.20	0.64	0.72	0.84	−2.20	0.64	0.72	0.84	−2.20	0.64	3.22	0.84
15	0.90	0.46	−0.40	−0.96	0.45	−0.54	0.19	−0.10	0.45	−0.54	0.19	−0.10	0.45	−0.54	2.69	−0.10
16	0.93	−0.52	0.32	−0.73	0.06	0.54	0.12	−0.72	0.06	0.54	0.12	−0.72	0.06	0.54	0.12	−0.72
17	1.34	−0.29	−0.11	−0.94	0.00	0.14	1.09	−1.23	0.00	0.14	1.09	−1.23	0.00	0.14	1.09	−1.23
18	1.64	0.21	−0.84	−1.01	−0.67	−0.21	1.42	−0.54	−0.67	−0.21	1.42	−0.54	−0.67	−0.21	1.42	−0.54
19	1.30	0.12	−0.62	−0.80	−0.78	−0.32	0.48	0.62	−0.78	−0.32	0.48	0.62	−0.78	−0.32	0.48	0.62
20	1.06	−0.29	−0.35	−0.42	−0.83	0.12	1.23	−0.52	−0.83	0.12	1.23	−0.52	−0.83	0.12	1.23	−0.52

$\sum_k^K \lambda_{ik} = 0$, $\sum_k^K \zeta_{gik} = 0$, and $\sum_g^G \pi_g = 1$, for identification, μ_θ and σ_θ set to zero and one for the first class. These priors, hyperpriors, and constraints were also used for analyzing the empirical data set in the later section. The computer software computer software WinBUGS (Spiegelhalter, Thomas, Best, & Lunn, 2007) was used for both simulation and empirical studies.

Convergence of the MCMC algorithm was examined using the Geweke test (1992) with a single chain as implemented in the computer program Convergence Diagnosis and Output Analysis for MCMC (CODA: Plummer, Best, Cowles, & Vines, 2006). Based on the Geweke test (1992) with a single chain and plots of autocorrelations, kernel density estimates of the marginal posterior distributions, and history plots of draws from posteriors, a burn-in of 3000 iterations was sufficient to achieve stationarity for all parameter estimates. Following this burn-in, an additional 5000 iterations were drawn to obtain estimates for each of the posterior distributions of model parameter estimates.

3.2 Simulation Study Results

3.2.0.1 Model Selection To investigate whether model fit indices could identify the correct number of latent classes, one- to four-class RIM-MixNRMs were fit to the data sets. The Bayesian information criterion (BIC: Schwarz 1978) and Akaike's information criterion (AIC: Akaike 1974) were chosen as model fit indices because of their popularity among researchers. Both indices were able to identify the correct model, that is, both indicated the three-class model was the best fit model for 93 of 100 replications. For the remaining replications, BIC suggested a two-class model and AIC a four-class model.

Label Switching Label switching is a well-known problem in finite mixture modeling. Two types of label switching can occur with mixture modeling (Cho, Cohen, & Kim, 2013). The first occurs over a single MCMC chain: the labels of the latent classes switch during estimation. The second type of label switching may be observed when labels switch between multiple data sets or multiple analyses in both Bayesian and maximum likelihood estimation. In the context of a simulation study, one needs to be aware of the possibility of label switching, as when labels switch on different replicate data sets, this may cause confusion when interpreting results. In the current study, the possibility of occurrence of label switching was investigated by inspecting profiles of item estimates across latent classes. When label switching was detected, latent classes were renamed by matching the profiles of parameter estimates across replications before calculating bias, mean squared error (MSE), and classification accuracy rates.

Recovery of Parameters A recovery analysis was done to determine whether the MCMC algorithm accurately recovered the model parameters of the RIM-MixNRM. In addition to inspection of label switching, parameter estimates had been placed on the metric of the generating parameters and then bias and MSEs

of parameter estimates were calculated. Correlations between generation values and estimates also used for the recovery analysis.

Results showed that most of the parameters of the RIM-MixNRM were recovered well: bias of item slopes, person ability and mixing proportion were about or less than 0.05; MSE were about or less than 0.16; and correlations were about or higher than 0.93. Correlations between item threshold parameters and estimates, ζ, was 0.94, but bias and MSE were -0.16 and 0.16, respectively, and appeared to depart slightly from generated values than other parameters. The RIM-MixNRM correctly classified examinees into their true (i.e., simulated) classes 87.85 % of the time.

4 Empirical Study: Instruction Effects on Students' Fractions Computation

In this section, we illustrate how a RIM-MixNRM can be used to help investigate effects of instruction on students' learning process. In this example, students' error patterns were examined on a test of fractions computation in a multi-year cluster randomized instructional intervention. The main purpose of the study was to investigate an experimental instructional condition effects on students' error patterns in computing fractions.

4.1 Data Description

Study Design The host study design was a school-based cluster randomized trial. Participants included 446 middle school students in Grades 6–8 in 25 general education math classrooms in 12 middle schools in and around a large metropolitan area in the Southeast. Students were randomly assigned to an experimental instructional condition (N = 214) or to a business-as-usual (BAU) condition (N = 232). There were 123 students with learning disabilities and 323 students without learning disabilities in the study.

The experimental condition implemented Enhanced Anchored Instruction (EAI; Bottge, Ma, Toland, Gassaway, & Butler, 2012). EAI was designed for use in helping to improve computation and problem solving skills of adolescents, including low-performing students with learning disabilities by including practical, hands-on applications to help students visualize the abstract concepts present in the problem. Teachers ask probing questions and offer instructional guidance to students as they view the video and help them identify relevant information to solve the problem. This eliminates the need for reading, a skill many low-achieving math students also lack.

Fractions Computation Test A Fractions Computation Test (FCT) consisting of 20 partial credit items (14-addition and 6-subtraction items) was designed by Bottge

et al. (2012) to measure students' ability to add and subtract simple fractions and mixed numbers with like and unlike denominators. The FCT was administered for the pre-test and the post-test. Math education experts identified 11 types of errors from students' incorrect responses to these items. The most common were *Combining* (*C*) and *Selecting Denominator* (*SD*). The remaining nine other types of errors occurred less frequently and were combined into a single *Other* (*O*) for this study. In the current study, the focus was on these three types of errors as they reflect students' misunderstandings about computing with fractions as well as the correct response (i.e., *No* errors). *Combining* and *SelectingDenominator* errors are described below:

- Combining (C): Student combines numerators and combines denominators, consistently applying the same operation to numerator and denominator.
- Select Denominator (SD): Student selects one of the denominators listed in the problem and makes no attempt to make equivalent fraction. Denominator given in the answer must be present in the problem.

4.2 Model Estimation

To investigate EAI effects on students' error patterns in fractions computation, we applied RIM-MixNRMs to take into account randomness in students and items parameters. The instructional method (i.e., EAI vs BAU) and students' math learning disability status (MD) were included in the model as covariates to predict the latent class membership as this could reflect of EAI effects on students' error patterns. This was done by substituting π_g in Eq. (3) with $\pi_{g|\mathbf{x}}$ as given by

$$\pi_{g|\mathbf{x}} = \frac{\exp(\beta_{g0} + \boldsymbol{\beta}_g' X)}{\sum_{g=1}^{G} \exp(\beta_{g0} + \boldsymbol{\beta}_g' X)} \tag{4}$$

where β_{g0} denotes an intercept for class g, and $\boldsymbol{\beta}_g = (\beta_{g1}, \beta_{g2}, \ldots, \beta_{gp})$ is a vector of logistic regression coefficients of covariates in the model. For this study, those covariates were the instructional method and students' math learning disability.

For those $\beta's$, normal distributions with mean of zero and standard deviation of 10 were used as conjugate priors and for rest parameters, the same priors were used as in the simulation study. For identification, the first class was used as a reference group.

The Geweke test, for convergence diagnosis, indicated that a burn-in of 4000 iterations was sufficient to achieve stationarity for all parameters. A subsequent 6000 iterations were used for estimating the model parameters. An exploratory analysis was applied to determine the number of latent classes in the data. That is, RIM-MixNRMs with from one- to five-class were fit to the pre- and post-test data. Based on BIC results, a four-class and a three-class RIM-MixNRMs for pre-test and post-test data, respectively were chosen for this study.

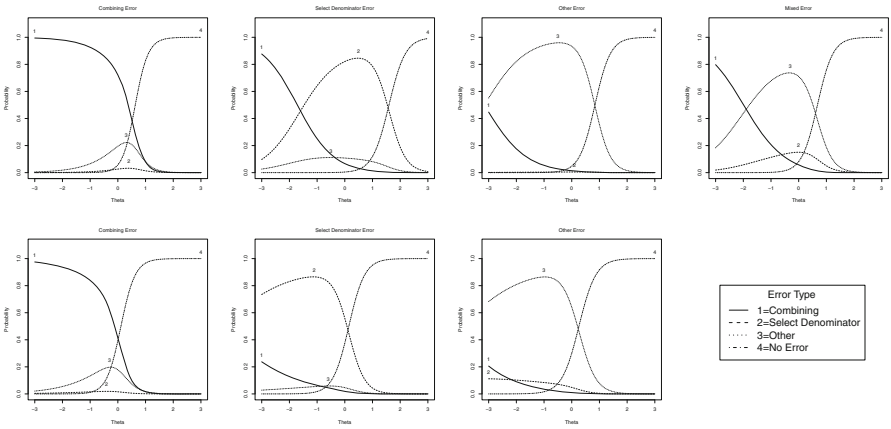

Fig. 1 Item category characteristic curves for item 16 showing latent class differences in students' error patterns on fractions computation

4.3 Results

4.3.1 Characteristics of Latent Classes

Item category characteristic curves (ICCC's) for Item 16 are shown in Fig. 1 for each latent class. These plots illustrate differences in types of errors made by students of the individual latent classes. The plots in the upper panel are for the pre-test and the plots in the lower panel are for the post-test. Students in all classes had a greater probability of not making any errors as they possessed more ability (i.e., Category 4). However, there were distinct error patterns which students in middle or lower ability in individual latent classes tended to make. Some students in middle or lower ability tended to mistakenly combine each numerator and each denominator (i.e., Category 1), some had a greater probability of making an error of selecting denominator (i.e., Category 2), some had a greater probability of making other errors (i.e., Category 3), and others had a greater probability of making either combining or selecting denominators (i.e., Category 1 or 2). Based on these patterns, each class is labeled as *Combining*, *SD*, *Other*, or *Mixed* shown in Fig. 1. These distinct differences can be interpreted as reflecting students' error pattern on the FCT.

4.3.2 Instruction Effects

Table 2 presents a cross-tabulation of the frequencies of students in each latent class on the pre- and the post-test. This shows a general pattern of students' transitioning in class membership from the pre-test to the post-test. To investigate effects of students' learning disability status and instructional type on such transitioning, those were included as covariates. On the pre-test, neither types of instruc-

Table 2 Transitioning pre-test to post-test latent classes

	Post-test			
Pre-test	Combine	SD	Other	Total
Combine	57	27	82	166
SD	3	25	33	61
Other	12	1	96	109
Mixed	19	16	75	110
Total	91	69	286	446

tion nor students' learning disability status did significantly impact on students membership in latent classes except that students with learning disability had significantly lower odds of belonging to *Other* error class than *Combining* error class (i.e., $\beta_{20} = -1.11, 0.33$ times). After the intervention, however, EAI had significant impact on students' error patterns on fraction computation. Students in EAI had significantly higher odds of belonging to *Other* or *SD* error classes than *Combining* error class (i.e., $\beta_{22} = 1.06, 2.89$ times and $\beta_{32} = 0.85, 2.34$ times, respectively). After the instruction, students might better understand about denominators and could distinguish them from numerators but still not fully understand the concept of common denominator in fractions.

5 Conclusion and Discussions

It is not uncommon that researchers or practitioners design an instrument to require nominal responses with a specific purpose in educational and psychological research area. For instance, in creating multiple choice items, item writers typically construct distractors in order to represent specific errors students might make. Nominal IRT models can be used to obtain information regarding these errors. Further, a mixture nominal IRT model can be used to take into account population heterogeneity; however, it does not consider randomness in items. In this study, we used a RIM-MixNRM in which both items and person parameters were considered a random sample from a population and taken into account their randomness. Results from a simulation study suggested that the model parameters were well recovered and both AIC and BIC provided useful information for model selection. Results from the middle school fractions computation data revealed there were four latent classes and three latent classes on the pre-test and the post-test, respectively, which could reflect students' error pattern on fractions computation. The results also show that instructional type had an significant impact on transitioning these error patterns subsequent to an instructional intervention. It is also possible to include item covariates. Inclusion of a Q-matrix (e.g., Tatsuoka 1983), for instance, as a covariate for individual categories of an item could be implemented to describe components of knowledge required for correctly answering a given question.

Acknowledgements The data used in the article were collected with the following support: the U.S. Department of Education, Institute of Education Sciences, PR Number H324A090179.

References

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

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 29–51.

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2001). A mixture model for multiple choice data. *Journal of Educational and Behavioral Statistics, 26*, 381–409.

Bottge, B., Ma, X., Toland, M., Gassaway, L., & Butler, M. (2012). *Effects of Enhanced Anchored Instruction on middle school students with disabilities in math*, Department of Early Childhood, Special Education, and Rehabilitation Counseling, University of Kentucky, Lexington, KY.

Cho, S.-J., Cohen, A. S., & Kim, S.-H. (2013). Markov Chain Monte Carlo estimation of a mixture item response theory model. *Journal of Statistical Computation and Simulation, 83*, 278–306.

De Boeck, P. (2008). Random item IRT models. *Psychometrika, 73*, 533–559.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian Statistics 4: Proceedings of the Fourth Valencia International Meeting* (pp. 169–194). Oxford: Oxford University Press.

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

Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). CODA: Convergence diagnosis and output analysis for MCMC. *R News, 6*(1), 7–11.

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

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

Spiegelhalter, D., Thomas, A., Best, N., & Lunn, D. (2007). WinBUGS, 1.4 [Computer program].

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*(4), 345–354.

Van den Noortgate, W., De Boeck, P., & Meulders, M. (2003). Cross-classification multilevel logistic models in psychometrics. *Journal of Educational and Behavioral Statistics, 28*(4), 369–386.

Wang, A. (2011). *A mixture cross-classification IRT model for test speededness*. Unpublished doctoral dissertation, University of Georgia.