교육평가연구 http://dx.doi.org/10.31158/JEEV.2018.31.3.707
Journal of Educational Evaluation
2018, 제31권 제3호, 707-728

Mixture IRT model for detecting aberrant responses under low-stake testing: Focusing on completely random guessing

Yoonsun Jang*
Korean Educational Development Institute

Georgia Wood Hodges**
University of Georgia

Item Response Theory (IRT) aids with the analysis of test items and item response patterns. One concern with test assessment is aberrant response. Aberrant responses are caused by several factors, for example, a participant's low level of motivation or their concern with the time limit for a high stakes test. These factors can yield biased estimates for IRT models that may invalidate inferences associated with research results. Mixture IRT models have been proposed and applied for detecting aberrant responses such as random guessing, performance decline in a low-stake test, or an extreme response style in a self-report survey. This study used a mixture Rasch model on an empirical data set to search for two latent classes defined as groups: valid responders or random responders. The effects of random responses on the estimation of model parameters was investigated as well as the effects of an educational intervention. In summary, only a small portion of random responders was detected. Consequently, they had a relatively weak effect on the estimation of model parameters and the efficacy of an educational intervention.

Keywords : item response theory, mixture IRT models, aberrant responses

* Korean Educational Development Institute, Researcher / Gyohak-ro 7, Ducksan-myeon Jincheon-gun, Chungcheongbuk-do, 27873 / Korea. 82-43-5309-263 / ysjang@kdei.re.kr

** University of Georgia, College of Education, Department of Mathematics and Science Education, Assistant Research Scientist / 110 Carton Street, Athens, GA 30602, / USA., 1-706-524-2257 / georgiahodges@uga.edu

Ⅰ. Introduction

An Item Response Theory (IRT) model estimates a person's latent trait based on their responses to the test or questionnaire items. One of the essential assumptions of IRT is that a person's responses to the items on an educational test or psychological questionnaire depend only on the latent trait that is being measured by the test(e.g., Baker & Kim, 2004). However, in reality this assumption can be violated due to a participant's level of motivation or level of fatigue, and so on. For example, when participants are not willing to expend their effort to respond to a test or survey, because the result of the test or survey is not very consequential to them, they may respond randomly or skip a question.

Although many researchers remove the cases that have a large portion(e.g., 50% or more) of missing responses, this strategy may not include the responders who answer all of the items but randomly guess(Guo, Rios, Haberman, Liu, Wang, & Paek, 2016). These unmotivated persons' random responses will not reflect their actual latent trait, consequently these particular responses can contribute to biased estimation of item and ability parameters. Another strategy to handle aberrant responses is to check person-fit statistics. Person-fit statistics measure the degree of goodness of fit for persons' responses to a statistical model. Although there are various profiles for person-fit statistics that are available, there was no universal agreement about a predominant one for identifying aberrant responses. The performances of person-fit statistics differed depending on the type of aberrant response, for example, carelessness, cheating, or random guessing as well as test length and percentage of aberrant responses(Karabatsos, 2003).

The application of mixture IRT(Rost, 1990) modeling has increased due to its ability to discern events such as aberrant responses. A mixture IRT model is a combination of a latent class model and a standard IRT model. Unlike the standard IRT model that assumes a homogeneous population, mixture IRT models detect different sets of item parameters within latent classes. Therefore, mixture IRT models provide a means for modeling heterogeneity in a population. The latent classes in mixture IRT models are unobserved and characterized based on patterns of item responses.

One major application for mixture IRT models is to detect different response patterns cause by

participant speededness or performance decline(e.g., Bolt, Cohen, & Wollack, 2002; Jin & Wang 2014b; Yamamoto & Everson, 1995). Speededness can occur when a person responds randomly to items at the end of a test due to an imposed time limit. If a test taker runs out of time on a test, this random(or guessing) strategy is often recommended for high-stake testing or any test with a time limit. In this instance, the latent trait is no longer being measured or assessed. Yamamoto and Everson(1995) proposed a HYBRID model with latent classes having different model structures. The mixture IRT models use the same parametric model but with different parameter vectors for each latent class. Consequently, HYBRID models can detect latent classes as the number of switching points where speededness occurs.

Bolt, Cohen, and Wollack(2002) also applied a mixture IRT model with ordinal constraints to detect speededness. Bolt, Cohen, and Wollack(2002) assumed that the item difficulty parameters at the end of test items were higher for the speededness class than those for the non-speededness class. Unlike Yamamoto and Everson(1995), however, Bolt, Cohen, and Wollack(2002) limited the number of latent classes to two: one is the speededness class and another is the non-speeded class. Namely, there is only one possible switching point. Jin and Wang(2014b) proposed a new model to control performance decline, which can be caused by speededness or a loss of motivation in a low-stake test. As with Yamamoto and Everson(1995), Jin and Wang(2014b) allowed varying switch points across persons. Jin and Wang(2014b) also allowed for the different probability of correct response based on each person's ability even after the switching point; whereas, the same probability was assumed for items after the switching point in Yamamoto and Everson(1995).

List, Robitzsch, Lüktke, Köller, and Nagy(2017) compared the performance of mixture models to detect a performance decline in a low-stake educational assessment. Three mixture models were applied to an empirical data set: (a) the Two-class Mixture Performance Decline Model(2PDM) proposed by Bolt, Cohen, and Wollack(2002); (b) the HYBRID model proposed by Yamamoto and Everson(1995); and (c) the Multiclass Mixture Performance Decline Model(MPDM) proposed by Jin and Wang(2014b) as well as a standard Two-Parameter Logistic Model(2PLM). List et al. (2017) found that all three mixture models fitted the empirical data better than 2PLM, but the proportions of persons who showed Performance Decline(PD) differed with each model. With

regard to the estimation of the model parameters, the item parameters estimated by three mixture models were relatively similar, but all three IRT models differed from those presented by the logistic model, 2PLM. Unlike to the estimated item parameters, 2PDM and 2PLM provided relatively similar ability parameters, while the ability parameters provided by HYBRID and MPDM were different from those presented by 2PLM.

Mixture IRT models were also applied to distinguish different response styles(e.g., extreme response style(ERS) and mild response style(MRS)) in Likert-scale data. Huang(2016) focused on distinguishing response styles in self-report survey data. For distinguishing three different response styles (a) normal, (b) MRS, and (c) ERS, Huang(2016) extended a ERS model with Generalized Partial Crediting Modeling(ERS-GPCM) suggested by Jin and Wang(2014a) to the mixture models, which is a mixture ERS-GPCM model. The results of simulation study in Huang(2016) showed that the model parameters estimated by ERS-GPCM were biased when the different response styles were not considered. Similarly, Jin, Chen, and Wang(2017) showed that inattentive responses in rating(or Likert) scale data caused seriously biased parameter estimation, and that the mixture GPCM model performed better than person-fit statistics *lz*(Drasgow, Levin, & McLaughlin, 1987) to detect attentive responses.

Previous research has shown that aberrant responses due to low motivation or speededness occur in test data; new models proposed for handling aberrant responses have discerned their effect better than the standard IRT models. Although new models are more appropriate statistically, adjusting a participant's score by compensating for aberrant responses can be controversial in a high-stake testing(Jin & Wang, 2014b). Thus, it is more appropriate for the new models to be applied by studies investigating group differences or exploring instrument trends rather than using these new models for estimating and reporting individual scores. This study focused on exploring the consequence of ignoring random responses in a low-stake test data set based on an empirical study. First, we used the mixture IRT model to detected random responses in an empirical data set collected from a low-stake test. Second, the effect of random responses on the model parameter estimation was investigated. Furthermore, we explored how random responses affect the evaluation of an educational intervention.

Ⅱ. Method

1. Detecting random responders

As described above, mixture IRT models are available to detect latent classes defined by a different set of item parameters for each latent class. This study applied a mixture Rasch model (MRM) to distinguish random responses from valid responses. MRM can be written as below;

$$P(u_{ij}=1) = \sum_{c=1}^{C}\pi_c \times P(u_{ij}=1|c,\theta_{jc}) = \sum_{c=1}^{C}\frac{\pi_c}{1+\exp(-\theta_{jc}+\beta_{ic})}, \quad (1)$$

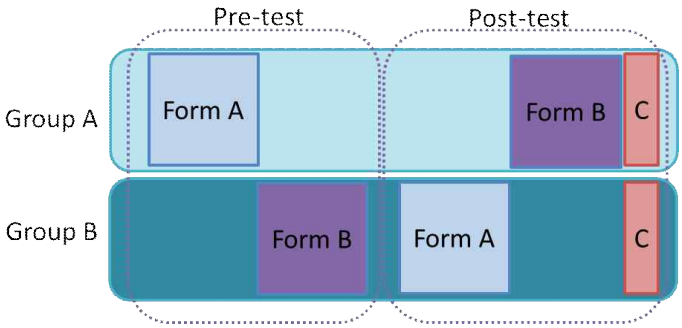
where u_{ij} is the response of person j to the item i, c represents a latent class, π_c is the probability of membership in latent class c, with the constraint $\sum_{c=1}^{C}\pi_c = 1$, θ_{jc} is the ability of person j who belongs to the latent class c, and β_{ic} is the item difficulty of item i for the latent class c.

Two latent classes were assumed: one class was a group of random responder; and the other class was a group of valid responders as proposed by Mislevy and Verhelst(1990). For the group of valid responders, the probability of correct response is defined by Equation 1, and all item parameters in Equation 1 are estimated. On the other hand, for the group of random responders, the same probability of correct response to the all items was assumed. Further, it was assumed that only the number of options following the item question were involved in calculating the probability of a correct response to that item. For example, the probability of correct response to a multiple choice item with five options is .20.

2. Data

The data set used in this study was taken from an article by Oliver, Hodges, Moore, Cohen, Jang, Brown, Kwon, Jeong, Raven, Jurkiewicz, and Robertson(2017), which examined the effect of a 3-D animation module on high school students' ability to learn biological concepts. Oliver et al.(2017) is a quasi-longitudinal study. There were two data waves: the Year 1 wave was the

control group that did not receive any educational intervention, the Year 2 wave consisted of data collected from focus groups where students who had received an educational intervention were questioned. We selected data from the Year 2 wave for this study. Three forms of assessment(Forms "A," "B," and "C") were developed to evaluate the effectiveness of the 3-D animations. Students were randomly assigned to two groups(Group "A" or "B"). At the beginning of semester, students who belong to Group A were given Form A as a pre-test, and students who belong to Group B were given Form B as a pre-test. After using the 3-D animations, students who belong to Group A were given Forms B and C as a post-test, and students who belong to Group B were given Forms A and C as a post-test. Thus, items in Form C were treated as anchor items for scaling of Form A and B. This data structure is described in [Figure 1].



〔Figure 1〕 Structure of empirical data

To maintain equivalency between Forms A and B, the substantive content of the items were kept, while only the question style of the items was changed. For example, in Form "A," item 1, the question is "The movement of water molecules across a selectively permeable membrane is called _____ ." Whereas in Form "B," item 1, the question is "Osmosis is defined as the diffusion of _____ ."

Forms A, B, and C contained 21, 21, and 8 items, respectively, and all items were multiple choice items with four options. In this study, the data consisted of the 627 students' responses to the pre- and (or) post-test.

3. Estimation of model parameters

In this study, the model parameters were estimated using the Markov chain Monte Carlo (MCMC) algorithm in the OpenBugs software(Spiegelhalter, Thomas, Best, & Lunn, 2014). Since we have no anchor items for scaling of the pre-test, the post-test data were used to estimate item parameters. Item parameters for Forms A, B, and C were calibrated simultaneously by using the items on Form C as the anchor items. After estimation of item parameters from the post-test data, item parameters for the pre-test data were fixed to the same values of estimated item parameters from the post-test data. Based on the assumption of estimation invariance across testing administration, by doing this, estimated parameters were put on the same scale, thus estimated student ability parameters from the post-test and pre-test data could be comparable.

The model parameters estimated by MRM for the post-test data were as follow: (a) the item difficulties(β_{ic}) of 50 items, which include the 21 items from Form A, the 21 items from Form B, and the 8 items from Form C, for a group of valid responders; (b) the ability parameters (θ_{jc}) of 587 students who took the post-test; and (c) the latent class memberships "c" of those students. The latent class membership indicates that each student is either a random responder or a valid responder, denoted with subscript 1 and 2, respectively. Additionally, the probabilities of membership in latent class(π_1 and π_2) and the means of ability parameters for two latent class (μ_{θ_1} and μ_{θ_2}) were estimated by MRM. It is important to note that there was no item difficulty estimated for the group of random responders because the random responders' probabilities of correct response to the all items were fixed at .25(i.e., 1/4), which corresponds to the probability of a correct answer for items with four options.

Based on previous research studies that applied MRM(e.g., Li, Cohen, Kim, & Cho, 2009; Sen, Cohen, & Kim, 2016), the prior and hyper prior distributions were specified as follow: (1) $c \sim Mltinomial(1, (\pi_1, \pi_2))$; (2) $(\pi_1, \pi_2) \sim Dirichlet(.5, .5)$; (3) $\theta_{jc} \sim Normal(\mu_{\theta_c}, 1)$; (4) $\mu_{\theta_c} \sim Normal(0, 1)$; and (5) $\beta_{i1} \sim Normal(0, 1)$. Again, it should be noted that the prior distribution of the item difficulty for the group of random responders was not specified because there was no item difficulty estimated for the class of random responders.

It is important to monitor convergence of Markov chains when an MCMC algorithm is used

to estimate model parameters. Convergence of a Markov chain is necessary in order to guarantee that drawing samples are from a stationary and representative distribution match the target distribution(Kim & Bolt, 2007; Sinharay, 2004). An MCMC algorithm is initiated with random samples, and these staring samples may or may not be close to the high density point of the posterior distribution(Sahlin, 2011). For this reason, samples drawn at the beginning of the chain are discarded. In this study, we discarded the samples drawn from the first 10,000 iterations and kept those from the next 10,000 iterations to avoid the effect of invalid samples. In addition to discarding the first 10,000 iterations, the convergence of all estimated item parameters was tested based on Heidelberger and Welch's(1983) method, which is one popular method to assess the stationary characteristic of a single chain.

To determine whether random responders existed in the empirical data, the model fit statistics of the Rasch model and MRM with two latent classes(the random responder and the valid responder) were compared. Two types of information indices, Akaike's(1973) an Information Criterion(AIC) and Bayesian Information Criteria(BIC)(Schwarz, 1978), were used to compare and select the most appropriate model. These two information criterion are the commonly used to find the best-fitting model in IRT(e.g., Choi & Wilson, 2015; De Boeck & Leuven, 2008; Li, et al., 2009; Sen, Cohen, & Kim., 2016). AIC and BIC are defined as below;

$$AIC = -2\log(L) + 2p, \tag{2}$$

$$BIC = -2\log(L) + p \times \log(N), \tag{3}$$

where L is the maximized value of the likelihood function, p is the number of estimated parameters, and N is the sample size. In this study, L was replaced by the posterior mean of the likelihood obtained via MCMC algorithm as described in Congdon(2003). The smaller value of model information criterion means the better model fit. Thus, the smaller AIC (or BIC) of the two-class MRM than those of the Rasch model suggests the existence of random responders.

Ⅲ. Results

The convergence of the estimated item difficulties was monitored by Heidelberger and Welch (1983), prior to the comparison of the results with the Rasch model and the two-class MRM (2C-MRM). Among 50 estimated item difficulties, 50(i.e., 100%) and 48(i.e., 96%) item difficulties were passed for Rasch model and 2C-MRM, respectively. That is, there was no convergence issue for both models. As described above, AIC and BIC were compared to select the best-fitting model for the pre- and post-tests. The AIC and BIC values are reported in <Table 1>. For the pre-test, AIC and BIC for the 2C-MRM were smaller than those of Rasch model regardless of the test form. Also, for the post-test, AIC and BIC for the 2C-MRM were smaller than those of Rasch model. That is, the 2C-MRM fitted better than Rasch model did to the empirical data. This result suggested that a group of random responders existed for both the pre-tests and the post-tests associated with the empirical data.

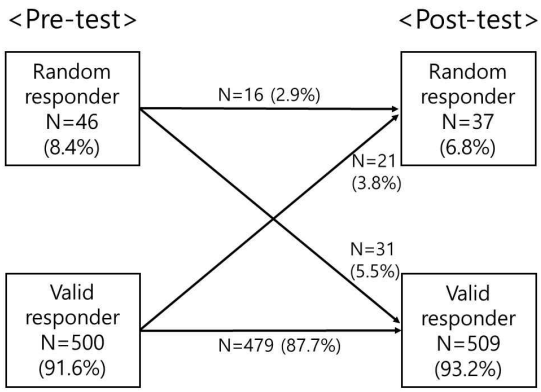
〈Table 1〉 Model information criterion indices for Rasch model and 2C-MRM

	Pre-test (Form A)		Pre-test (Form B)		Post-test (Forms A, B & C)	
	Rasch	2C-MRM	Rasch	2C-MRM	Rasch	2C-MRM
AIC	7,594.00	7,550.00	7,166.00	7,088.00	30,580.00	30,086.00
BIC	7,597.70	7,557.41	7,169.66	7,095.31	30,798.75	30,317.88

<Table 2> shows the number of students who were detected as a random responder or a valid responder. The percentage of random responders ranged between 6.75% and 10.14%. Overall, the percentage of random responders decreased from the pre-test to the post-test for both test forms. The percentage of random responders for Form B was somewhat larger than those for Form A for the both pre- and post-tests. Among 587 students who took the post-test, 546 students took the pre-test, as well. {Figure 2} indicates how many students moved to the random responder at the pre-test from the valid responder at the post-test, or vice versa. The percentages of each transition were: (a) the percentage of the random to random was 2.9%, (b) the percentage of the random to valid was 5.5%, (c) the percentage of the valid to random was 3.8%, and (d) the percentage of the valid to valid was 87.7%.

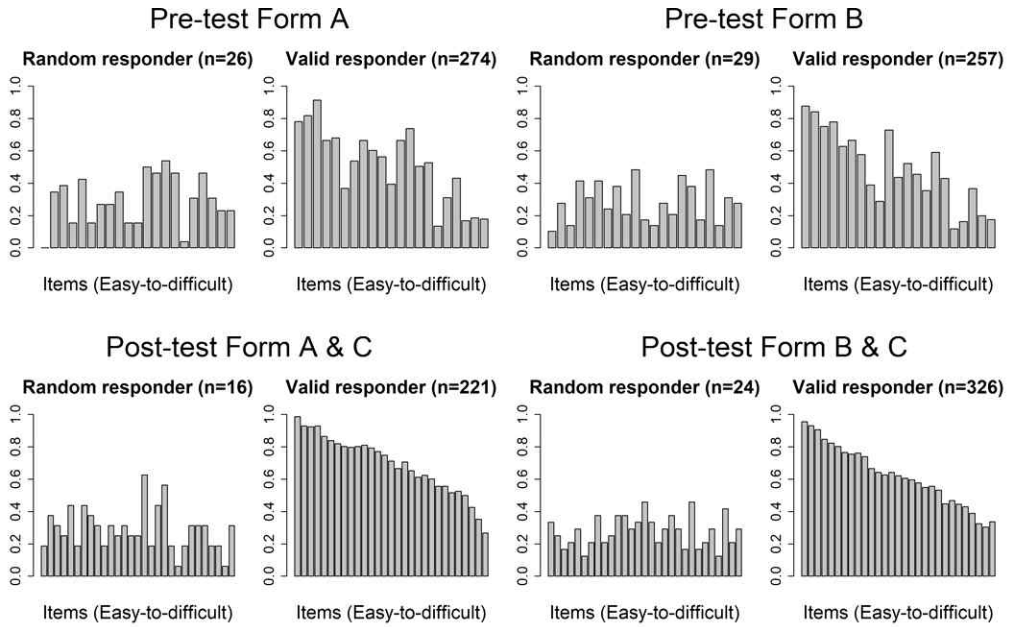
〈Table 2〉 The number of students who were assigned to each latent class by 2C-MRM

	Pre-test				Post-test			
	Form A		Form B		Forms A & C		Forms B & C	
	N	%	N	%	N	%	N	%
Random responder	26	8.667	29	10.140	16	6.751	24	6.857
Valid responder	274	91.333	257	89.860	221	93.249	326	93.143
Total	300	100	286	100	237	100	350	100



〔Figure 2〕 The percentages of latent class transitions

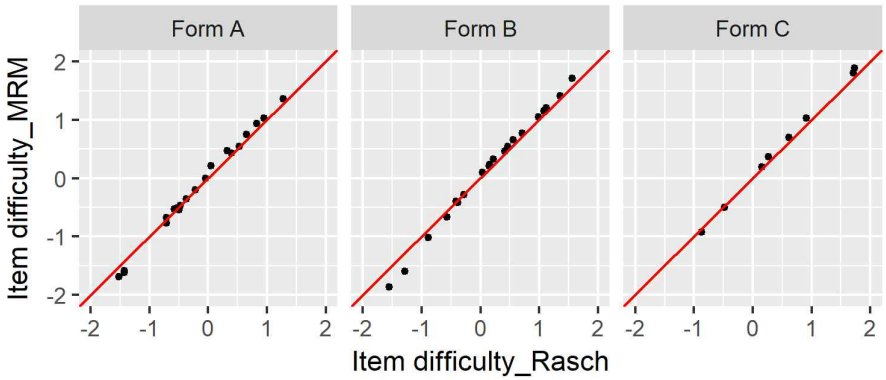
The proportions of correct responses for the classes of random responders and valid responders are compared in [Figure 3]. There are four sets of plots based on the test form(Form A or B) and the time point(the pre- or post-test). Each set has two plots: one is for the class of random responders, and another is for the class of the valid responders. The x-axis of plots in [Figure 3] represents items ordered from easy to difficult(21 items for the pre-test and 29 items for the post-test), and the y-axis represents the proportion of correct response. The item difficulties used to order the items were estimated by the Rasch model with post-test data, because the item difficulties for the pre-test were fixed at the same values as the item difficulties that were estimated based on the post-test. Thus, the item difficulties for the two test forms were on the same scale, as noted above.



〔Figure 3〕 Patterns of proportions of correct response to the items ordered by easy to difficult for the groups of random responders and valid responders.

As expected, there was no specific pattern of proportions of correct response for the group of random responders regardless of the test forms and time points. For the group of valid responders, however, the proportions of correct response showed a decreasing trend from the left to right of x-axis(that is, as the items became difficult). This decreasing trend for the post-test was more clear than those for the pre-test. To sum up, the Rasch model did not fit to the item responses for the group of random responders. To explore the effect of random responders on the estimation of item and ability parameter, the item and ability parameters estimated by Rasch model were compared with those by the 2C-MRM.

Plots in 〔Figure 4〕 display the relations between the item difficulties estimated by Rasch model and 2C-MRM for the items of test forms A, B, and C, respectively. The x-axis of each plot represents the item difficulties estimated by Rasch model, and the y-axis of each plot represents the item difficulties estimated by 2C-MRM. As can be seen in 〔Figure 4〕, the item difficulties estimated by Rasch model and 2C-MRM were close to each other regardless of the test forms. The correlations between item difficulties estimated by Rasch model and those by 2C-MRM were



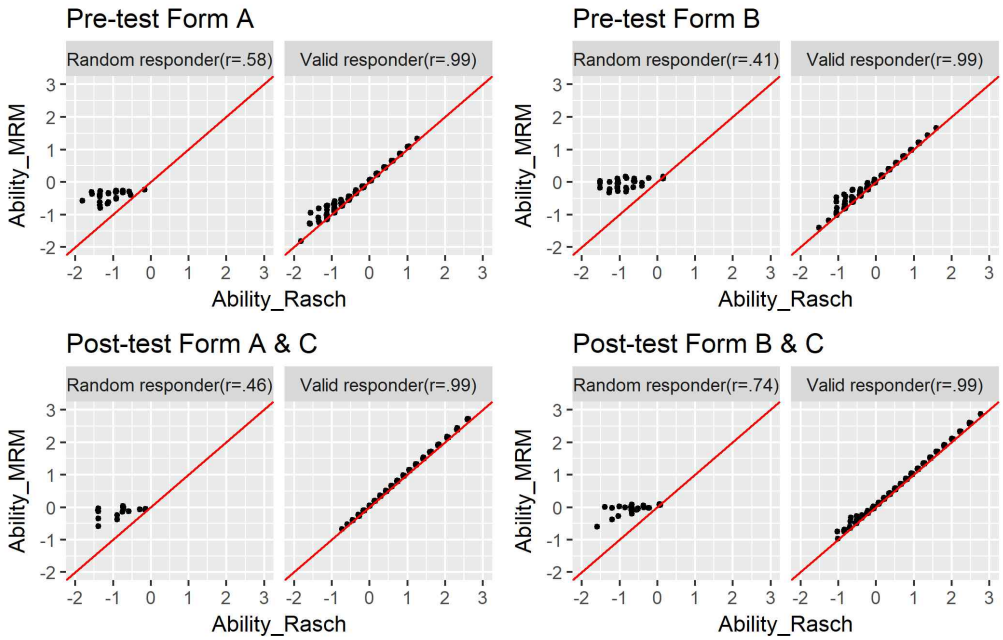
〔Figure 4〕 Relations between item difficulties estimated by Rasch model and those by 2C-MRM for each test form

above .99 for the all three test forms. The item difficulties estimated by the Rasch model were slightly larger than those by the 2C-MRM, as the items became easy. On the other hand, as the items became difficult, the item difficulties estimated by the Rasch model were slightly smaller than those by the 2C-MRM. That is, the item difficulties estimated by Rasch model tended to be shrunken, this is consistent with the results from Jin, Chen, and Wang(2017).

The relations between ability parameters estimated the by Rasch model and those by the 2C-MRM are showed in [Figure 5]. There are four sets of scatter plots as in [Figure 3]. Similar to the plots in [Figure 4], the x-axis represents the ability parameter estimated by Rasch model, and the y-axis represents the ability parameter estimated by 2C-MRM.

As can be seen in [Figure 5], for the random responders, the ability parameters estimated by the Rasch model scattered between -2 and 0, whereas those by the 2C-MRM were located near to 0 regardless of the test forms and the time points. The correlations between the ability parameters estimated by the Rasch model and those by the 2C-MRM model were .58, .41, .46, and .74, for the pre-test Forms A, B and the post-test Forms A, B, respectively. The plausible reason of the small variance of the ability parameters estimated by the MRM for the random responders is that the prior distribution of ability parameters was specified as a standard normal distribution, and the probability of correct answer was fixed at .25 for all random responders.

For the valid responders, the ability parameters estimated by the 2C-MRM were almost identical to those by the Rasch model, particularly for the post-test. The correlations between



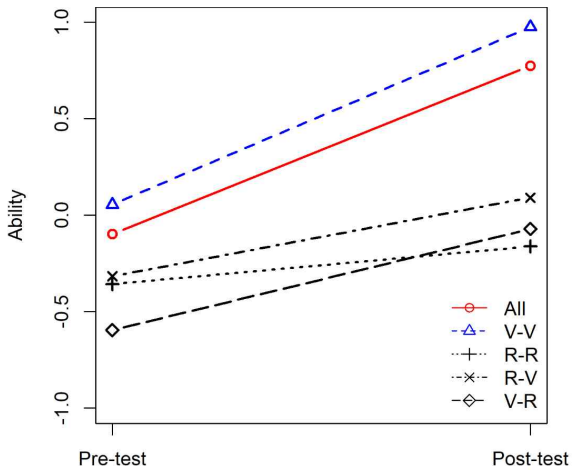
〔Figure 5〕 Relations between ability parameters estimated by Rasch model and those by 2C-MRM for the groups of random responders and valid responders.

ability parameters estimated by the Rasch model and those by the 2C-MRM were above .99 for all time points and test forms. The ability parameters estimated by Rasch model for the pre-test were smaller than those by 2C-MRM when the estimated ability parameters were below zero. That is, Rasch model tended to underestimate ability parameters particularly for the examinees who had relatively low ability. For the post-test, the ability parameters estimated by the Rasch model and MRM were very close to each other across all ranges of the ability scale, although the estimated ability parameters for the post-test Form B were slightly dispersed at below zero.

A paired sample t-test was conducted to explore the influence of random responders associated the 3-D animation module course. First, a paired sample t-test was used with all responders(i.e., ability parameters estimated by the Rasch model), and then a paired sample t-test was used with only valid responders(i.e., ability parameters estimated by 2C-MRM for the group of random responders), at the both pre- and post-test stages. The paired sample t-test with all responders indicated that the estimated ability parameters were significantly higher for the post-test than for the pre-test, $t(332) = -24.58$, $p = .000$, $dz = 1.35$. After removing random responders, the

paired sample t-test also showed that the difference of ability parameters at the pre- and post-test was statistically significant, $t(297) = -25.56$, $p = .000$, $dz = 1.48$. That is, the difference of estimated ability parameters between the pre- and post-test was statistically significant, whether removing the random responders or not. However, the effect size, a standardized Cohen's d(dz), was slightly increased, from 1.35 to 1.48 when the random responders were removed at both the pre- and post-test.

[Figure 6] shows the means of estimated ability parameters at the pre- and post-test, with respect to the groups of random responders and valid responders. As can be seen in [Figure 6], the means of estimated ability parameters at the pre- and post-test grew from -0.01 and 0.06 to 0.78 and 0.98 when random responders were removed(see All represented by the solid line and V-V represented by the short-dash line in [Figure 6]). Moreover, the means of two groups of students who were distinguished as random responders(see R-R represented by the dotted line and V-R represented by the long-dash line in [Figure 6]) were smaller than those of students who were distinguished as valid responders(see R-V represented by the dot-dash line and V-V represented by the short-dash line in [Figure 6]) at the post-test. At the pre-test, however, the mean of V-R, a group of students who were distinguished as valid responders at the pre-test, was lower than that of R-R and R-V, which are groups of examinees who were distinguished as random responders. The lower mean of V-R than those of R-R and R-V was result from the



[Figure 6] Group means of estimated ability parameters at the pre- and post-test. V means a valid responder and R means a random responder.

prior distribution of ability parameters for a class of random responders.

Ⅳ. Discussion

The pre- and post-test research design is common in educational and psychological studies that examine the efficacy of an intervention. It is also common that a persons' responses are affected by many factors like a time limit, low motivation, and so on, which are called nuisance factors. These nuisance factors cause several types of aberrant responses including performance decline and random responses. Without handling aberrant responses in data, the validity of research result can be threaten. The main goals of this study were to demonstrate how random responses affect the model parameters estimation and consequently have influence on investigating the effect of an educational intervention.

We applied the Rasch model and the 2C-MRM to the empirical data from Oliver et al. (2017). Oliver et al.(2017) administrated pre- and post-tests to examine the effects of a 3-D animation module that was developed to improve the learning of biological concepts for high school students. Because the pre- and post-test conducted in Oliver et al.(2017) did not provide important benefits(e.g., a high school diploma or a license), it could be regarded as a low-stake test. Based on the results of this study, AIC and BIC preferred 2C-MRM to Rasch model. The results supported the ability of new models to detect random responses in the empirical data, furthermore, this level of discernment helps to guarantee valid testing results.

The proportions of the random responders detected by 2C-MRM were relatively small, ranged between about 7% and 10% at the pre- and post-test. Fortunately, the effect of random responses on investigating educational intervention(i.e., 3-D animation module) was not strong. These results suggested that the most students who were involved in Oliver et al.(2017) might be highly motivated and responded carefully because this study carried out by their school teachers during school hours. Although the random responses did not result in misleading the effect of 3-D animation on students' learning, the result of this study showed shrinkage of the estimated item parameters. Additionally, the ability parameters were underestimated due to the shrinkage of item parameters, particularly at the below zero point of the ability scale. That is,

the random responders may cause biased ability parameters for students who are located on the extreme ends of ability continuum. Consequently, the inference based on examinees' ability parameters estimated without removing random responders can be biased.

In this study, we mainly focused on detecting completely random guessing in a low-stake test situation. For this purpose, the probability of correct response was fixed at .25, which is simply computed by 1/c, with c is the number of options, for all items of the test because the probability of correct response does not depend on item properties, such as item content the level of item difficulty, or the degree of attractiveness of options(Han, 2012). However, logical guessing is also possible when an examinee has partial knowledge. The examinee with partial knowledge can remove some unattractive options or is be more likely to select an attractive incorrect option. In the former case, the probability of correct response of this examinee would be lager than 1/c, and vice versa in the latter case. Since logical guessing also can affect the estimation of model parameters, some researchers may want to detect random guessing as well as logical guessing. In this case, estimating the probability of correct response might be more appropriate than fixing to a certain value for a latent class of guessing responders.

Although this study mainly focused on detecting random guessing, we additionally analyzed our empirical data using the 2C-MRM without fixing probability of correct response of the random responder class. The estimated probability of correct response of the random responders was about 32.7%, and AIC and BIC of the 2C-MRM with estimating probability for the post-test were 30,108 and 30,344.25, respectively. AIC and BIC were larger than those of the 2C-MRM with fixing probability(AIC is 30,086 and BIC is 30,317.88 as reported in Table 1). That is, the 2C-MRM with fixing probability fitted the empirical data better than with estimating probability.

Mixture IRT models can be used in either an exploratory or a confirmatory approach. When mixture IRT models used in an exploratory approach, determining the number of latent classes and defining the characteristics of latent classes extracted by a mixture IRT model are critical issues. To focus on the main purpose of this study, the mixture IRT model was utilized in a confirmatory approach as in Jin, Chen, and Wang(2017). That is, we assumed two latent classes (random responders and valid responders) and did not consider additional latent classes. Although speededness is common in a high-stake test and the empirical data were not from a high-stake test, however, performance decline caused by speededness might have occurred. An exploratory

mixture IRT analysis might be useful to detect several types of aberrant responses. For example, the number and characteristics of latent classes determined by exploratory mixture IRT analysis can provide evidence of additional latent class, such as speededness. An additional latent class for a performance decline can incorporate simultaneously and easily to MRM, that is, MRM with three latent classes(valid responders, random responders, and performance decliners). In this sense, the 3C-MRM can provide better understanding of empirical data.

As shown by the results of this study and other previous studies(e.g., Bolt, Cohen, & Wollack, 2002; Osborne & Blanchard, 2011), the mean of ability parameters for a group of examinees who had aberrant responses like random guessing or speededness tended to be lower than that of valid responders. These results were reasonable, but might not be true because we do not know the true ability of examinees with aberrant responses. Simulation study to explore how well MRM can distinguish the random responders and examinees with low level of ability would be helpful to understand the performance of MRM for detecting random responders. Moreover, simulation study with multiple conditions that can consider various kinds of aberrant response patterns would be needed for better understanding of the performance of MRM.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csáki (Eds.), *2nd International Symposium on Information Theory* (pp. 267-281). Budapest, Hungary: Akadémiai Kiadó.

Baker, F. N., & Kim, S. -H. (2004). *Item response theory: Parameter estimation techniques.* New York, NY: Marcel Dekker, Inc.

Bolt, M. D., Cohen, A. S., & Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: Application of a mixture Rasch model with ordinal constraints. *Journal of Educational Measurement, 39,* 331-348.

Choi, I. -H., & Wilson, M. (2015). Multidimensional classification of examinees using the mixture random weights linear logistic test model. *Educational Psychological Measurement, 75,* 78-101.

Congdon, P. (2003). *Applied Bayesian modelling.* New York, NY: John Wiley.

De Boeck, P., & Leuven, K. U. (2008). Random item IRT models. *Psychometrika, 73,* 533-559.

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1987). Detecting in appropriate test scores with optimal and practical appropriateness indices. ***Applied Psychological Measurement, 11,*** 59-79.

Guo, H., Rios, J. A., Haberman, S., Liu, O. L., Wang, J., & Paek, I. (2016). A new procedure for detection of students' rapid guessing responses using response time. ***Applied Measurement in Education, 29,*** 173-183.

Han, K. T. (2012). Fixing c parameter in the three-parameter logistic model. ***Practical Assessment, Research & Evaluation, 17***, 1-24.

Heidelberger, P., & Welch, P. (1983). Simulation run length control in the presence of an initial transient. ***Operations Research, 31,*** 1109-1144.

Huang, H. -Y. (2016). Mixture random-effect IRT models for controlling extreme response style on rating scale. ***Frontier in Psychology, 7***: 1706.

Jin, K. -Y., Chen, H. -F., & Wang, W. -C. (2017). Mixture item response models for inattentive responding behavior. ***Organizational Research Methods, 21,*** 197-225.

Jin, K. -U., & Wang, W. -C. (2014a). Generalized IRT models for extreme response style. ***Educational and Psychological Measurement, 74,*** 116-138.

Jin, K. -U., & Wang, W. -C. (2014b). Item response theory models for performance decline during testing, ***Journal of Educational Measurement, 51,*** 178-200.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. ***Applied Measurement in Education, 16,*** 277-298.

Kim, J. -S., & Bolt, D. M. (2007). Estimating item response theory models using Markov chain Monte Carlo methods. ***Educational Measurement: Issues and Practice,*** 38-51.

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

List, M. K., Robitzsch, A., Lüktke, O., Köller, O., and Nagy, G. (2017). Performance decline in low-stakes educational assessment: different mixture modeling approaches. ***Large-scale Assess in Education, 5,*** 15.

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

Oliver, J. S., Hodges, G. W., Moore, J. N., Cohen, A., Jang, Y., Brown, S., A., Kwon, K. A., Jeong, S., Raven, S. A., Jurkiewicz, M., & Robertson, T. P. (2017). Supporting high school student accomplishment of biology content using interactive computer-based curricular case studies. ***Research in Science Education,*** 1-26.

Osborne, J. W., & Blanchard, M. R. (2011). Random responding from participants is a threat to the valid of social science research results. ***Frontiers in Psychology, 1,*** 220.

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

Sahlin, K. (2011). ***Estimating convergence of Markov chain Monte Carlo simulations*** (Master's thesis). Retrieved from http://www2.math.su.se/mastat/ reports/master/2011/rep2/report.pdf

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

Sen, S., Cohen, A. S., & Kim, S. －H. (2016). The impact of non-normality on extraction of spurious latent classes in MixIRT models. ***Applied Psychological Measurement, 40,*** 98-113.

Sinharay, S. (2004). Experiences with Markov chain Monte Carlo convergence assessment in two psychometric examples. ***Journal of Educational and Behavioral Statistics, 29,*** 461-488.

Spiegelhalter, D., Thomas, A., Best, N., & Lunn, D. J. (2014). ***OpenBUGS User Manual***, Version 3.2.3. MRC Biostatistics Unit, Institute of Public Health and Department of Epidemiology and Public Health, Imperial College School of Medicine, UK, available at http://www.mrc-bsu.cam.ac.uk/bugs.

Yamamoto, K., & Everson, H. T. (1995). ***Modelling the mixture of IRT and pattern responses by a modified HYBTID model*** (Report No. RR-95-16). Princeton, NJ: Educational Testing Service.

◎ 논문접수: 2018. 7. 16 / 수정본 접수: 2018. 9. 8 / 게재승인: 2018. 9. 14

저 자 소 개

· 장윤선 : University of Georgia에서 교육측정(Quantitative Methodology) 전공으로 박사 학위 취득. 현재 한국교육개발원 대학역량진단센터 위촉 연구원. 관심 연구 분야는 문항반응이론, 준거설정, 잠재집단모형 등임. ysjang@kedi.re.kr

· Georgia Wood Hodges : University of Georgia에서 과학 교육 전공으로 박사 학위 취득. 현재 University of Georgia 수학·과학 교육학과 연구교수. 관심 연구 분야는 과학 교사 연구, 과학 교육용 게임 및 학습 테크놀로지를 통한 학습 유도 등임. georgiahodges@uga.edu

〈국문요약〉

혼합 문항반응이론 모형을 적용한 저부담 검사 환경에서의 비정상적인 문항반응 추출 - 완전 무선 추측을 중심으로

장 윤 선
한국교육개발원

Georgia Wood Hodges
조지아대학교

문항반응이론(IRT) 모형은 검사 문항 및 검사 문항에 대한 반응유형 분석 등에 사용된다. 피험자의 문항반응은 검사에서 측정하고자 잠재특성 외 다른 요인들에 의해 영향 받으며 이로 인하여 비정상적인 문항반응이 발생하기도 한다. 예를 들어, 피험자가 검사에 흥미를 가지지 못한 경우 문항에 성실히 응답하지 않을 수 있으며, 고부담 검사에 참여한 피험자가 제한된 시간에 대한 불안으로 비정상적인 문항반응을 보일 수 있다. 비정상적인 문항반응은 문항반응이론 모형의 편파적인 모수 추정을 야기할 수 있으며, 이로 인해 왜곡된 연구 결과가 도출 될 수 있다. 본 연구에서는 혼합 문항반응이론(mixture IRT) 모형을 적용하여 비정상적인 문항반응을 추출하고, 추출된 비정상적인 문항반응이 문항반응이론 모형의 모수 추정에 미치는 영향을 탐색하였다. 아울러 교육적 처치의 효과 분석에 비정상적인 문항반응이 미치는 영향을 분석하였다. 분석 결과, 추출된 비정상적인 문항반응의 비율은 비교적 낮았으며, 이들이 교육적 처치 효과 분석에 미치는 영향 또한 크지 않았다.

주제어 : 문항반응이론, 혼합 문항반응이론 모형, 비정상적 문항반응

Appendix

OpenBUGS code used for the 2c-MRM

```
## Class 1 = random responder, Class 2=Rasch model
## NI =  number of items, NE = number of examinees
## b2 = centered item difficulty, beta2 = item difficulty
## gmem2 = class membership, pi2 = mixing proportion
## theta2 = ability parameter, mut2 = mean of ability parameters
model
{
for (j in 1:NE) {
 for (k in 1:NI) {
   tt2[j,k]<- exp(theta2[j] - b2[gmem2[j],k])*a2[gmem2[j],k]
   p2[j,k]<-tt2[j,k]/(1 + tt2[j,k])+c2[gmem2[j],k]
   resp[j,k]~dbern(p2[j,k])
   l2[j,k]<-log(p2[j,k])*resp[j,k]+log(1-p2[j,k])*(1-resp[j,k])
}

## priors for ability
 theta2[j] ~ dnorm(mut2[gmem2[j]],1)
 gmem2[j] ~ dcat(pi2[1:G2])
}
for (j in 1:G2) {
 mut2[j]~ dnorm(0,1)
}
pi2[1:G2]~ ddirch(alpha2[])
## priors for item difficulty of Class2
for (j in 2:G2){
 for (k in 1:NI){
```

```
    beta2[j,k]~dnorm(0,1)
    b2[j,k]<-beta2[j,k]-mean(beta2[j,1:NI])
   }
  }
  ## fixed parameters for Class1
  for(k in 1:NI){
   a2[1,k]<-0
   a2[2,k]<-1
   b2[1,k]<-0
   beta2[1,k]<-0
   c2[1,k]<-0.25
   c2[2,k]<-0
  }
  loglik[2]<-sum(l2[1:NE,1:NI])
  }

  list(NE=587, NI=50,G2=2, alpha2=c(.5,.5),
  resp = structure(.Data = c(
  1,0,1,1,1,1,1,0,1,1,1,0,0,1,1,1,0,1,0,1,1,NA,NA,NA,NA,NA,NA,NA,
   ⋮
  0,1,1,0,1,1,0,0,0,0,1,0,0,0,1,0,0,0,0,0,0,1,0,0,0,1,0,0,0),
  .Dim=c(587,50)))
```