Psychological Test and Assessment Modeling, Volume 64, 2022 (1), 65-86

Comparison of Different Approaches to Dealing with Guessing in Rasch Modeling

Hong Jiao[1]

Abstract

This study compared three approaches to dealing with guessing in Rasch modeling: explicit modeling of guessing effects, correction of guessing effects, and the Rasch model which does not model guessing explicitly. The extended Rasch model explicitly includes a lower asymptote parameter in the Rasch model to account for guessing. Parameter estimation was explored using a Bayesian approach for the extended Rasch model with guessing. Further, model parameter estimates were compared with those from the Rasch model and the Rasch model with the correction procedure for guessing effects under different study conditions. The results indicated that the true model parameters could be well recovered by the Bayesian estimation method developed in OpenBUGS. Ignoring guessing in general led to the overestimation of test information, underestimation of item difficulty, and misrepresentation of the maximum test information location.

Keywords: The Rasch model; guessing; test information; MCMC Bayesian estimation

[1] University of Maryland, College Park, USA *Correspondence concerning this article should be addressed to:* Hong Jiao, Measurement, Statistics and Evaluation, Department of Human Development and Quantitative Methodology, 1230C Benjamin Building, University of Maryland, College Park, MD 20742, USA, hjiao@umd.edu

The Rasch model (Rasch, 1960) is widely applied in test or instrument development and item response data analysis. In the Rasch model, the probability of a correct response to an item is related to an examinee's latent ability and one item characteristic, item difficulty. Setting constraints on item parameters including uniform item discrimination (1), zero low asymptote (no guessing), and unity upper asymptote (no slipping), the Rasch model possesses some unique characteristics such as sufficiency, separability, and consistency, different from other item response theory (IRT) models that other non-Rasch models do not possess. More specifically, raw scores are sufficient statistics for person parameter estimation. As highlighted by one reviewer, according to Anderson (1973) and Fischer (1974), only in the Rasch model, both model parameters can be estimated by the conditional maximum likelihood (CML) estimation method (Scheiblechner, 2009). However, in real testing situations, aberrant item response behaviors such as guessing and slipping may be present due to a variety of reasons. Low ability examinees may attain a correct response to an item whose difficulty is above their ability level due to guessing (McDonald, 1967, p. 67). In the Rasch modeling framework, guessing is treated as an unexpected response. It is either eliminated from model parameter estimation (Linacre, 2000) or utilized to identify misfit persons (Smith, 1993; Wright, 1991). Artner (2016) compared five fit indexes in detecting person misfit in the Rasch model. Guessing is one of the responding behaviors simulated for misfit.

Guessing may occur as random guessing or smart guessing. Random guessing makes use of no prior information or information from the test and blindly chooses a response to an item (Roger, 1999). An example is in a speeded test, examinees may randomly select a choice due to running out of time (Wise & Demars, 2005). If examinees lack of motivation in taking the test or the items are too difficult, they may randomly select an option as well. Assuming randomly guessed items are multiple-choice items, examinees with low ability may still have the chance of guessing the item correct, $1/m$, where m is the number of options. Other times, examinees may guess smartly based on partial knowledge or synthesizing information from other sources such as prior knowledge or information in the item like wording cues, cues in item stems or distractors or other items on the test to remove least attractive distractors and increase their chance of a correct response (Lord, 1983; McDonald, 1989; Roger, 1999). In general, guessing affects the ability parameter estimation (Dinero & Haertel, 1977; van de Vijver, 1986) and item difficulty parameter estimation (Dinero & Haertel, 1977; Pelton, 2002).

Researchers explored different approaches to correcting (e.g., Choppin, 1983; Linacre, 2008) or modeling different types of guessing effects (e.g., Barton & Lord, 1981; Birnbaum, 1968; Cao & Stokes, 2008; Keats, 1974; San Martin, Del Pino, & De Boeck, 2006 ; Weitzmen, 1996). One method corrects the guessing effect by setting the pseudo-guessing parameter to a fixed value (most often an inverse of the number of options) for all items in the Rasch model (Barnes & Wise, 1991; Divgi, 1984; Smith & Fujimoto, 2011; Wainer & Wright, 1980). All these studies found that the Rasch model with fixed lower asymptote increases the ability parameter estimation accuracy compared with the Rasch model ignoring the guessing effect. Another correction

procedure is the CUTLO correction procedure implemented in WINSTEPS (Linacre, 2008) that eliminates an examinee's item response when the examinee's ability estimate is lower than the item difficulty estimate with a certain logit unit defined by CUTLO, so that the response will not contribute to the estimation of item difficulty parameter (Detail of the CUTLO procedure can be found in the WINSTEPS manual or at http://winsteps.com/winman/cutlo.htm). This procedure is consistent with Waller's ability removing random guessing model (e.g., Waller, 1973; 1989) and Choppin's procedure (1983).

Modeling guessing has been approached from two perspectives: one treats guessing as a psychometric property of an item while the other treats guessing as a person characteristics and guessers and non-guessers are assumed from different latent populations. When guessing is modeled as an item property, the common approach is to include a lower asymptote parameter in the item response theory (IRT) model such as Keats' generalized Rasch model for guessing (Keats, 1974), the three-parameter logistic (3PL) IRT model (Birnbaum, 1968) and the four-parameter logistic (4PL) IRT model (Barton & Lord, 1981). Wise and DeMars (2006) developed the effort-moderated model. If an examinee's response time is longer than the threshold, the model reduces to the 3PL IRT model. Otherwise, the model reduces to a constant probability model with the reciprocal of the number of response options as the guessing probability. Some other researchers (Nedelsky, 1954; San Martin, Del Pino, & De Boeck, 2006; Thissen & Steinberg, 1984) view that guessing is associated with an examinee's ability, and proposed ability-based guessing models. Other researchers (Hessen, 2004; 2005) consider that guessing depends on item difficulty and reparameterize item response theory models as constant latent odds-ratio models.

Cao and Stokes (2008) developed three mixture IRT models to accommodate different types of guessing behaviors by grouping guessers and non-guessers into different latent classes. One of their models assumes that examinees respond depending on their ability up to a certain item, and guess thereafter. An item location threshold is estimated for each examinee, indicating the item number at which guessing starts. This piecewise formulation of the item response modeling applies to the speeded test scenarios (Yamamoto, 1995). Another model assumes that examinees respond to easy items based on their ability and guess randomly on difficult items. This is a testwise skill often recommended to examinees to maximize their test performance. The third model accommodates guessing due to low motivation. It assumes that the guessers will make decreasing effort as they proceed through the test. These models classify each examinee into guesser and non-guesser classes and measure the degree of guessing behavior.

It is worthy of note that the inclusion of a lower asymptote parameter in the Rasch model is not limited to modeling the guessing effect. A non-zero lower asymptote in IRT modeling could model other pseudo-guessing effects which may lead to spuriously high scores when examinees correctly respond to difficult items which are above their ability levels. This may occur when examinees engage in cheating, answer copying, or know the correct answers to some items due to item disclosure (Chen & Jiao,

2012). In psychological tests, examinees may fake responses due to social desirability which will lead to a lower asymptote larger than zero.

This study focuses on one modeling approach, the extended Rasch model for guessing, which explicitly includes a lower asymptote parameter in the Rasch model to account for the guessing or pseudo-guessing effect. This Rasch model plus a lower asymptote does not any longer possess the unique characteristics of the Rasch model as described above and the CML estimation method no longer works. This explicit approach follows the conceptualization and parameterization of the guessing or pseudo-guessing effect in the standard non-Rasch IRT modeling framework though other modeling approach is possible. It explores the model parameter estimation for the extended Rasch model for guessing (Kubinger & Draxler, 2007; Linacre, 2002) using a Bayesian approach. Further a simulation study and a real data analysis are conducted to compare this explicit modeling approach with a correction approach which is the same as the CUTLO correction procedure in WINSTEPS and the Rasch model in terms of item difficulty and ability parameter estimation.

The Extended Rasch Model for Guessing

The Rasch model includes one item parameter and one latent ability parameter to describe the probability of a correct response to an item as follows.

$$P\ \ (x_{ij}|b_i, \theta_j) = \frac{1}{1+exp(-(\theta_j - b_i))}. \quad (1)$$

Keats (1974) introduced a constant guessing parameter as a lower asymptote into the Rasch model to take account of the guessing behavior. The constant guessing parameter is equal to the reciprocals of the number of options in the multiple-choice items. This approach has been utilized in other extended Rasch model such as the multiplicative Rasch model (Smith & Fujimoto, 2011) to account for guessing. Keats' generalized Rasch model for guessing is mathematically represented as in equation 2.

$$P\ \ (x_{ij}|b_i, c\ \ , \theta_j) = c\ \ + \frac{1-c}{1+exp(-(\theta_j - b_i))}. \quad (2)$$

Or alternatively, $log(\frac{P_{ij}-c}{1-P_{ij}}) = \theta_j - b_i$, where $P\ \ (x_{ij}|b_i, c\ \ , \theta_j)$ represents the conditional probability of a correct response for examinee *j* with ability θ_j to item *i* with item difficulty b_i and guessing parameter c , constant across all items. In educational tests, the presence of the guessing effect may be related to the non-zero probability of getting an item correct when a person's ability is asymptotically getting to $-\infty$. In psychological tests, the parameter may represent the probability of endorsing an item when a person's possession of the latent trait asymptotically goes to $-\infty$. Keats' generalization of the Rasch model for guessing retains the additivity property of the standard Rasch model (Keats, 1974), that is, the item and person parameters are separable into additive components (White, 1976). Though there is no consistent

maximum likelihood estimates (Colonius, 1977), Keats' generalization of the Rasch model for guessing have the sufficient statistics property when the guessing parameters are constant across all items (Linacre, 2002). Linacre (2002) developed the maximum likelihood estimator for the ability parameter for this quasi-Rasch model which allows varying guessing values for different items as presented in equation 3 and later developed an approximation of the guessing parameter based on the parameter estimates from an initial analysis using the Rasch model (Linacre, 2004).

$$P\ \ (x_{ij}\big|b_i, c_i, \theta_j) = c_i + \frac{1-c_i}{1+exp(-(\theta_j - b_i))}. \qquad (3)$$

As illustrated above, the idea of the Rasch model with guessing parameters is not new; several researchers (e.g., Keats, 1974; Kubinger, 2005; Linacre, 2002, 2004; Weitzman, 1996) explored adding a guessing parameter in the Rasch model to accommodate the guessing effect. As reported in Kubinger and Draxler (2007), Puchhammer (1989) explored the joint maximum likelihood estimation of the Rasch model for guessing. As expected, the joint maximum likelihood estimates are inconsistent for the number of fixed items. The item difficulty parameters are biased and the guessing parameter estimates are not accurate when the number of examinees is fewer than 500. Kubinger and Draxler (2007) explored constraining the standard 3PL IRT model in BILOG MG 3 (Zimowski et al., 2003) to estimate the parameters for the Rasch model for guessing using the marginal maximum likelihood estimation method. They focused on the fit comparison between the Rasch model and the Rasch model for guessing. They concluded the use of the Rasch model for guessing could save more items into the item pool as the Rasch model for guessing provided better fit. This current study explored a Bayesian approach to estimate model parameters for the extended Rasch model for guessing as presented in equation 3 (Linacre, 2002). Further, model parameter estimates were compared with those from the Rasch model which does not explicitly model the guessing effect and the Rasch model with guessing correction which is the same as the CUTLO procedure implemented in WINSTEPS under different study conditions.

Method

To investigate the model parameter recovery for the extended Rasch model for guessing, both simulation data and real data were analyzed. In the simulation study, both item difficulty parameters and ability parameters were simulated from a standard normal distribution with a mean of 0 and standard deviation of 1. Two sample sizes were specified for the ability parameters: 500 and 1,000. A sample size of 500 is considered as an adequate sample size while a sample size of 1,000 is considered as the recommended sample size for the Rasch model. The true difficulty parameters for the simulated 40 items remained the same across study conditions and replications while the ability parameters for the conditions with the same number of persons remained the same. The magnitude of guessing was manipulated at three levels, 0.1, 0.2, and 0.3; the same values were assigned to all items in each study condition. Smith (2008) simulated item response data with pseudo-guessing effects paired with item difficulty, that is, easy items have lower pseudo-guessing effects and difficult items have higher pseudo-guessing effects. This is one special case for all possible pseudo-guessing effects. In real data analyses, the correlation between item difficulty and pseudo-guessing parameters could be very low (Jiao, Macready, Zhu, & An, 2011; Kubinger & Draxler, 2007) when guessing is included as a lower asymptote parameter in the IRT models. Difficult items may have lower pseudo-guessing effects when students lack of motivation to take the test. On the other hand, easy items may have higher pseudo-guessing effects or spuriously higher probability of a correct response due to copying and item disclosure. Thus, the use of uniform values for the pseudo-guessing parameters helps to better investigate the impact of the guessing effects.

Item responses were generated based on equation 3 using the true model parameters. Item response data were analyzed with six procedures including the extended Rasch model for guessing, the Rasch model, and the Rasch model with the CUTLOW correction by specifying four CUTLOW values: 0.5, 1, 1.5, and 2 (that is, if a person's ability is 0.5 logit unit lower than the item difficulty and this person correctly responded to the item, the correct response will be considered as due to potential guessing and recoded as missing and will not contribute to model parameter estimation). Though Smith (2008) did not find significant impact of guessing on ability parameter estimation when comparing the Rasch model estimates and the estimates from the CUTLOW procedure in WINSTEPS (Linacre, 2008), he did not include the true model in the comparison. Thus, this current study explored the true model estimation and compared the estimation errors from the true models and the alternatives. By fully crossing the levels of the magnitudes of guessing, sample sizes, and the methods in dealing with the guessing effect, thirty-six study conditions were simulated.

This study developed a Markov Chain Monte Carlo (MCMC) estimation algorithm in OpenBUGS 3.2.1 (Lunn, Thomas, Best, & Spiegelhalter, 2000) for the extended Rasch model for guessing. The priors for the ability parameter followed a standard normal distribution with a mean of 0 and standard deviation of 1. The priors for the item difficulty parameters were set normally distributed with a mean of 0 and variance

of 2. The larger variance served as a relatively less informative but proper prior. The priors for the guessing parameter followed a beta distribution with the first shape parameter α and the second shape parameter β specified as follows. When the guessing effect was simulated at 0.1, the beta distribution was specified with a α of 3 and a β of 19 to obtain a mode of 0.1 (Baker & Kim, 2004). Beta distributions were specified with a α of 5 and a β of 17 to obtain a mode of 0.2, and with a α of 7 and a β of 15 to obtain a mode of 0.3 respectively (Baker & Kim, 2004). This informative prior for the guessing parameters was to remove the potential source of error due to the misspecification of the priors for the model parameters in the MCMC estimation. The Rasch model and the Rasch model with the CUTLO correction procedures were also implemented in OpenBUGS to remove potential differences due to different estimation programs.

The MCMC iterative algorithm ran two Markov chains in parallel, each starting with different initial values supplied by OpenBUGS. Convergence was checked based on multiple criteria. The Gelman-Rubin statistic as modified by Brooks and Gelman (1998) was used. Convergence can be assumed if R < 1.05 (Lunn et al., 2000). A sample check over replications in the study conditions indicated that R was generally close to 1 and smaller than 1.05 before 40,000 iterations. The Brooks-Gelman Ratio (BGR) diagnostic plots and the trace plots indicated that stability and convergence usually occurred between 30,000 and 40,000 iterations. The quantile plots showing the running mean with 95% confidence intervals against iteration numbers indicated that the running mean and the 95% confidence intervals from the two chains mixed very well and reached equilibrium before 40,000 iterations. Other plots including history and density plots all indicated that the two chains mixed well before 40,000 iterations and reached equilibrium by then. Thus, the first 40,000 iterations were discarded as the burn-in iterations. An additional 10,000 iterations were monitored for each chain. The model parameter inferences were made based on a total of 20,000 samples.

Simulation for each study condition was replicated twenty times to compute estimation errors in item and ability parameters in terms of bias, standard error (SE), and root mean squared error (RMSE) for each of the thirty-six study conditions. The bias, SE, and RMSE were computed based on equations 4, 5, and 6 respectively.

$$Bias(\hat{\beta}) = \frac{\sum_{r=1}^{N}(\hat{\beta}_r - \beta)}{N}, \tag{4}$$

$$SE(\hat{\beta}) = \sqrt{\frac{1}{N}\sum_{r=1}^{N}\left(\hat{\beta}_r - \bar{\hat{\beta}}\right)^2}\ , \tag{5}$$

$$RMSE(\hat{\beta}) = \sqrt{\frac{1}{N}\sum_{r=1}^{N}(\hat{\beta}_r - \beta)^2}, \tag{6}$$

where β is the true model parameter, $\hat{\beta}_r$ is the estimated model parameters for the r^{th} replication, $\bar{\hat{\beta}}_r$ is the average of the estimated model parameters over r replications, and N is the number of replications. The average bias, SE, and RMSE are computed by averaging each of the values over all item or ability parameters.

One real data set from a large-scale science test was analyzed. The analysis of the real data was not intended to select a better fitting model, rather to better understand the differences in model parameter estimates among the studied procedures. Further, test information and test characteristic curves were compared as well. Test information for the Rasch model for guessing was derived and is presented in equation 7.

$$I_T(\theta) = \sum \quad I_i(\theta) = \sum \frac{(P_i^{'}(\theta))^2}{P_i(\theta)Q_i(\theta)} = \sum \frac{Q_i(\theta)(P_i(\theta)-c_i)^2}{(1-c_i)^2 P_i(\theta)}, \tag{7}$$

where $I_i(\theta)$ is the item information function for a specific theta point, $P_i(\theta)$ is given in equation 3, $Q_i(\theta) = 1 - P_i(\theta)$, and c_i is the pseudo-guessing parameter for item i. Test characteristic curve function is given in equation 8.

$$T\ (\theta) = \sum \quad P_i(\theta). \tag{8}$$

Results

Simulation Data

Only significant effects at a significance level of 0.05 with at least small effect sizes are reported. Non-significant or significant results with negligible effect sizes are not reported. The magnitude of effect size is classified as negligible (f<0.1), small (0.1<f<0.25), moderate (0.25<f<0.4), and large (f>0.4) in the analysis of variance.

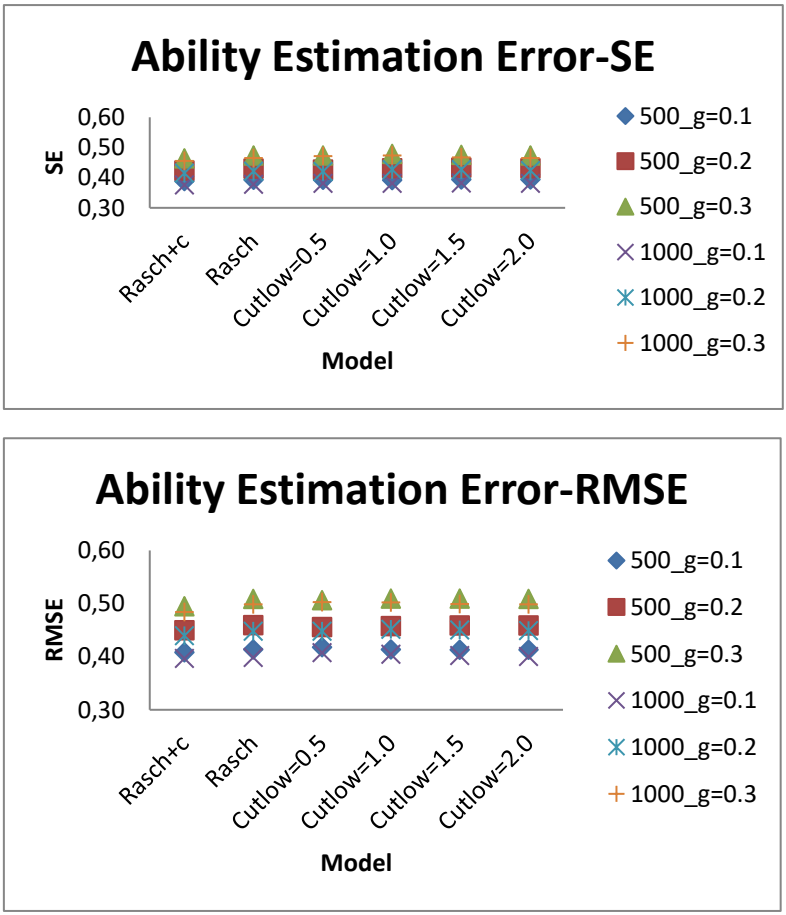

***Figure 1*. SE and RMSE in the ability parameter estimates.**
Note: Two sample sizes: 500 and 1000; three guessing effects: g=0.1, 0.2, and 0.3.

Ability Parameter Estimation. To assure scale identification within a model and the comparability of the model parameter estimates from different models and methods, ability parameter estimates were rescaled and item parameter estimates were adjusted accordingly. Thus, there was no difference in the bias for ability estimates. Two sample sizes were simulated. The estimation errors in the ability parameters are summarized in Figure 1. The patterns in the estimation errors were similar for both sample sizes. Only the results for the sample size of 500 are reported here. Only the guessing magnitude significantly affected the SE and RMSE in the ability parameter estimation with large effect sizes (f=0.40 and f=0.41 respectively). Post-hoc Tukey analyses indicated that all pairwise contrasts for SE and RMSE among different guessing magnitudes were significant. Both SE and RMSE in the ability parameter estimation increased as the guessing effect increased.

Item Difficulty Parameter Estimation. The estimation model, guessing magnitude and their interaction all significantly impacted the bias in the item difficulty estimation with large effects (f=3.28, f=3.45 and f=1.53 respectively). The mean biases are presented in Figure 2 and the interaction between the model and the guessing magnitude is presented in Figure 3. The ordinal interaction generally indicated that the Rasch model with a lower asymptote produced the least bias in the item difficulty estimation. The Rasch model with a CUTLOW correction value of 0.5 produced the second least bias while the Rasch model without any correction led to the largest bias. The Rasch model with a correction value of 2 did not effectively reduce the impact of guessing because only a few correct item responses were recoded as missing due to the large correction value. The Tukey procedure indicated that all pairwise differences among different guessing magnitudes were significant and all pairwise differences among the six estimation models were significant except that between the Rasch model and the Rasch model with a CUTLOW correction value of 2.0.

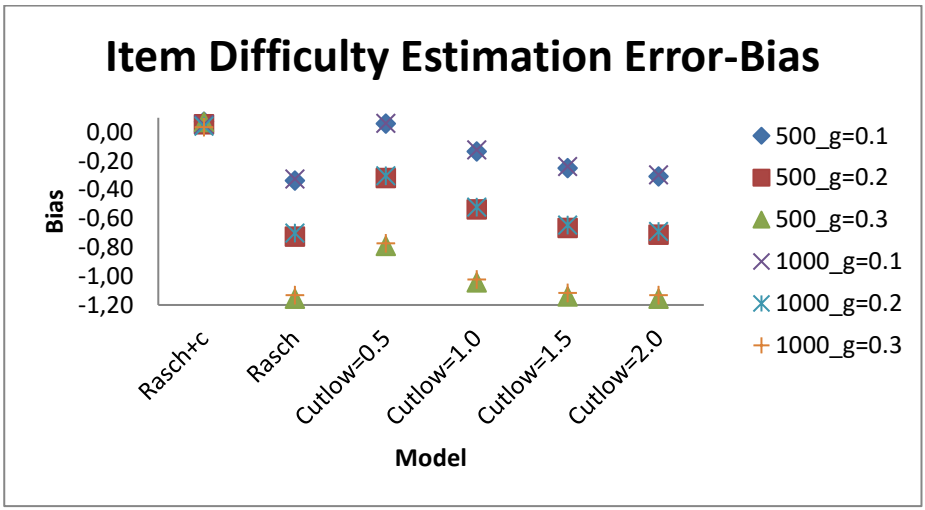

***Figure 2*. Bias in the item difficulty parameter estimates.**

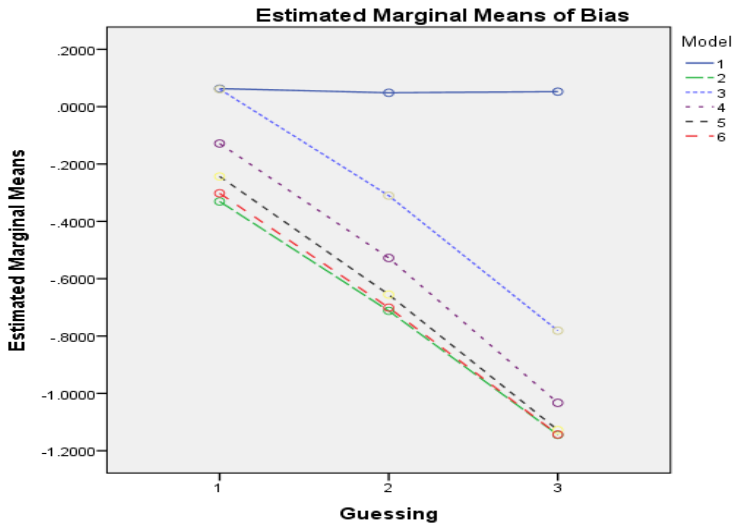

***Figure 3*. The effect of the interaction between the estimation model and the guessing magnitude on the bias in the item difficulty parameter estimates.**
Notes: Model 1=Rasch with a lower asymptote, Model 2=Rasch model, Model 3, 4, 5, and 6=Rasch model with a CUTLOW value of 0.5, 1.0, 1.5, and 2.0 respectively; Guessing 1, 2, and 3=lower asymptotes of 0.1, 0.2, and 0.3.

In general, ignoring the guessing effect in Rasch modeling leads to underestimation of item difficulty. It is expected that the larger CUTLOW values led to a smaller number of examinees deleted. As the guessing effect increased, the number of examinees retained for item calibration became larger. This is counterintuitive since it is expected that larger guessing effects should lead to more misfit responses. However, a possible explanation for this can be provided based on Figure 4 which presents the item characteristic curves for four items with the same difficulty but different guessing effects. Given the same ability level, the probability of a correct response will be the highest for the item with the highest guessing effect. On the other hand, given the same probability of a correct response or expected item score, the difficulty would be the lowest for the item with the highest guessing effect when calibrated with the Rasch model as the ICC is going to shift to the left. Or when the guessing effect is ignored, the item would be estimated to be easier.

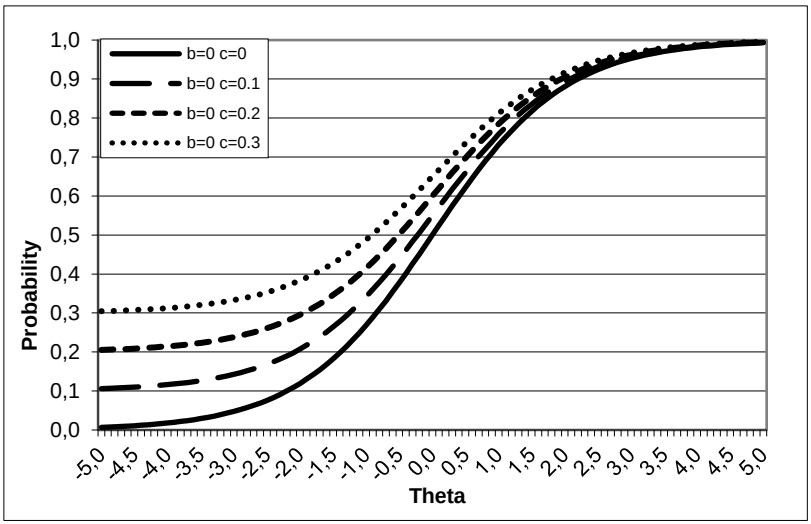

***Figure 4*. Item characteristic curves for items with the same difficulty but different magnitudes of guessing effects.**

All three studied factors, model, sample size, and guessing magnitude significantly impacted the SE in the item difficulty estimation (see Figure 5) with large effect sizes (f=0.60, f=0.74, and f=0.42 respectively). The SE was also significantly influenced by the interaction between model and the guessing magnitude (see Figure 6) with a small effect size (f=0.21). In general, the increase in the guessing magnitude increased the random error in the item difficulty parameter estimation. The ordinal interaction between the model and the guessing magnitude indicated that the Rasch model produced the smallest random error while the Rasch model with a lower asymptote for guessing

yielded the highest random errors. A possible explanation is that the Rasch model is a simpler model with fewer parameters while the Rasch model with guessing has more parameters to be estimated; thus, the increase in the number of parameters in the Rasch model for guessing increased the random error in the item difficulty parameter estimation. When trimming the item response data with the CUTLOW procedures, the random error was affected by the extent of missing data. The smaller CUTLOW value increased the amount of missing item responses, which lead to higher random error. The Tukey pairwise comparison indicated that all pairwise SE differences among different guessing magnitudes were significant and only the pairwise differences between the Rasch model with lower asymptotes and each of the other models were significant. The pairwise differences among the Rasch model and the Rasch models with different correction values were not significant.

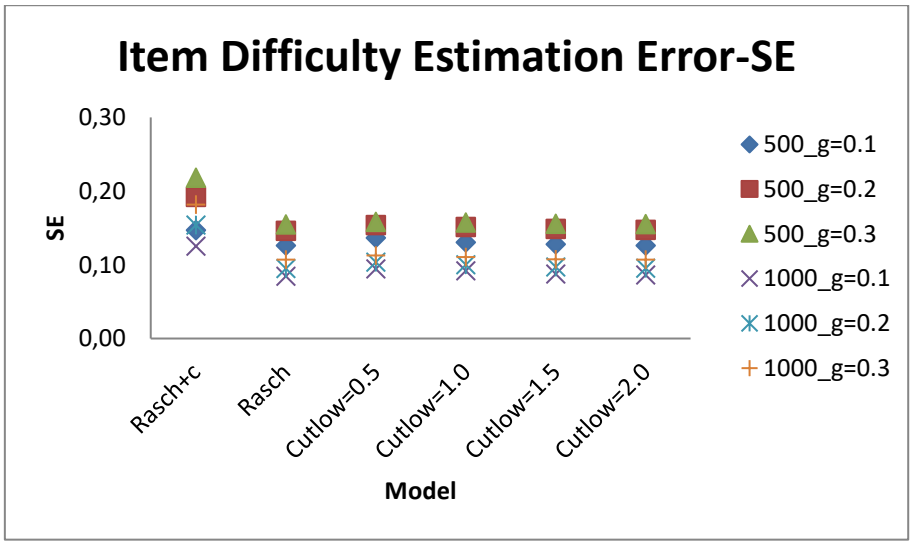

***Figure 5*. SE in the item difficulty parameter estimates.**

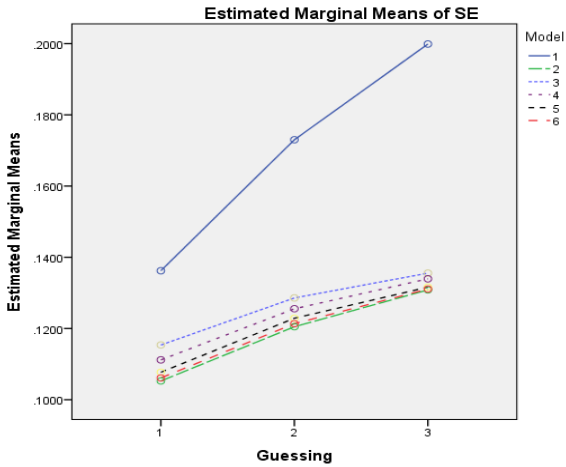

***Figure 6*. The effect of the interaction between the estimation model and the guessing magnitude on the SE in the item difficulty parameter estimates.**
Notes: Model 1=Rasch with a lower asymptote, Model 2=Rasch model, Model 3, 4, 5, and 6=Rasch model with a CUTLOW value of 0.5, 1.0, 1.5, and 2.0 respectively; Guessing 1, 2, and 3=lower asymptotes of 0.1, 0.2, and 0.3.

Model, guessing magnitude and their interaction significantly impacted the total error in the item difficulty estimation with large effect sizes (f=2.58, f=3.60, and f=1.54). The effect of the sample size was small (f=0.20). In general, the larger guessing effects increased the total errors, RMSE (see Figure 7). The ordinal interaction between the model and the guessing magnitude (see Figure 8) indicated that the Rasch model with lower asymptotes produced least RSME while the Rasch model yielded about the largest RMSE. Post-hoc contrasts with the Tukey's procedure indicated significant pairwise differences among different guessing magnitudes and significant pairwise differences among different models except the difference between that the Rasch model and the Rasch model with a CUTLOW correction value of 2.0 which could be considered as a non-effective correction value. Overall, the removal of potential misfit item responses purifies the item response data thus decreases the total error in the item difficulty parameter estimation.

Guessing Parameter Estimation. The bias in the guessing parameter estimation for the Rasch model for guessing was not affected by either the magnitude of guessing or the sample size. The random error in the guessing parameter estimation was significantly affected by the guessing magnitude with a large effect size (f=0.59). As the guessing effect increased, the random error increased as well. Sample size had a small effect on the random error (f=0.15). The increase in the sample size decreased the random error. Both sample size and the magnitude of guessing had small effects on the total estimation error in the guessing parameter (f=0.22, f=0.16). The increase in the sample size decreased the total error and the increase in the guessing magnitude increased the total error in the guessing parameter estimation.

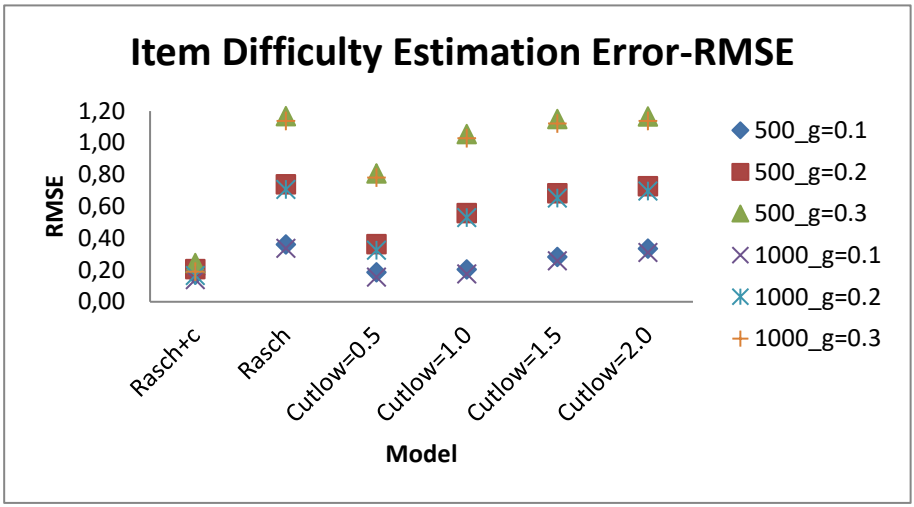

***Figure 7*. RMSE in the item difficulty parameter estimates.**

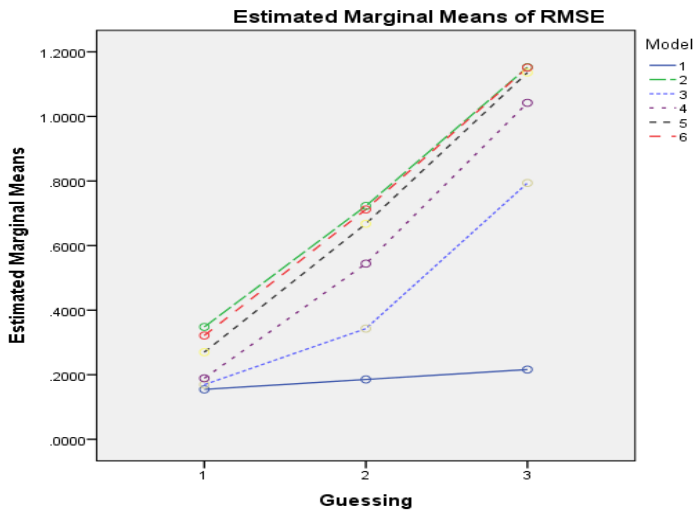

***Figure 8*. The effect of the interaction between the estimation model and the guessing magnitude on the RMSE in the item difficulty parameter estimates.**
Notes: Model 1=Rasch with a lower asymptote, Model 2=Rasch model, Model 3, 4, 5, and 6=Rasch model with a CUTLOW value of 0.5, 1.0, 1.5, and 2.0 respectively; Guessing 1, 2, and 3=lower asymptotes of 0.1, 0.2, and 0.3. '

Real Data

A large-scale science test was analyzed with each of the compared procedures. The test consisted of 40 items. Item responses from 623 examinees were available for analysis. Since no true values of the guessing effects were known, multiple priors for the guessing parameter were explored to select a prior with better fit. Like in the simulation study, the same beta distributions were specified to obtain a mean guessing effect of 0.1, 0.2, and 0.3. In addition, a beta distribution was specified with a αvalue of 4 and a β value of 18 to obtain guessing effects with a mode of 0.15 and with a α value of 6 and a β value of 16 to obtain a mode of 0.25 (Baker & Kim, 2004). Four fit indices were used to select the prior with better fit: Akaike's information criterion (AIC; Akaike, 1974), the Bayesian information criterion (BIC; Schwarz, 1978), a version of the AIC corrected for small sample sizes, AICc (Sugiura, 1978) and Deviance Information Criterion (DIC; Spiegelhalter et al., 2002). The fit indices are presented in Table 1. In general, all fit indices except AIC supported that the beta prior with a mode of 0.1 provided the best fit.

Table 1: ***Fit Indices for the Real Data with Different Prior Distributions***

Mode of Guessing	0.1	0.15	0.2	0.25	0.3
AIC	26430	26430	26440	26460	26470
AICc	26450	26460	26470	26480	26500
BIC	26780	26790	26800	26810	26830
DIC	26840	26850	26850	26860	26880

Ability parameter estimates are summarized in Table 2. Since the scale comparability was achieved by standardizing the ability parameter estimates, there was no difference in the mean and standard deviation of the ability estimates across methods. The correlations among the ability estimates across methods were almost all above 0.99. In general, items were estimated to be more difficult by the Rasch model plus guessing than by other five methods (see Table 2). The correlations between item difficulty estimates of the Rasch model plus guessing and other models ranged from 0.93 to .95 while the difficulty estimation among the Rasch model and the Rasch model with correction were over 0.99. The correction procedure with a CUTLOW value of 0.5 yielded item difficulty estimates closest to those from the Rasch model with lower asymptotes. The correction procedure with a CUTLOW value of 2 essentially produced the same estimates as the Rasch model.

Test information curves are presented in Figure 9. The ability point with the maximum test information differed. The Rasch model for guessing was around 0.8 while the others were around -0.8. The test information was much lower for the Rasch model for guessing than other models. A further examination of similar simulation conditions with a

sample size of 500 and guessing of 0.1 supported the findings. This indicates that the ignorance of the guessing effect lead to overestimation of the test information along the ability scale and misrepresentation of the maximum test information location.

Table 2: ***A Summary of the Ability and Item Difficulty Parameter Estimates for the Real Data***

Ability	N	Minimum	Maximum	Mean	Standard Deviation
Rasch+c	623	-2.7828	2.3445	0.0000	1.0000
Rasch	623	-3.0782	2.5612	0.0000	1.0000
CUTLO=0.5	623	-2.8721	2.1456	0.0000	1.0000
CUTLO=1.0	623	-3.2182	2.3025	0.0000	1.0000
CUTLO=1.5	623	-3.2832	2.4512	0.0000	1.0000
CUTLO=2.0	623	-3.3983	2.5307	0.0000	1.0000
Item Difficulty	N	Minimum	Maximum	Mean	Standard Deviation
Rasch+c	40	-1.7546	1.7473	-0.3484	0.9413
Rasch	40	-2.2270	0.6044	-0.8648	0.7792
CUTLO=0.5	40	-1.8559	1.0346	-0.5480	0.8044
CUTLO=1.0	40	-2.0006	0.8451	-0.6892	0.7880
CUTLO=1.5	40	-2.1232	0.7443	-0.7952	0.7769
CUTLO=2.0	40	-2.1898	0.6629	-0.8416	0.7769

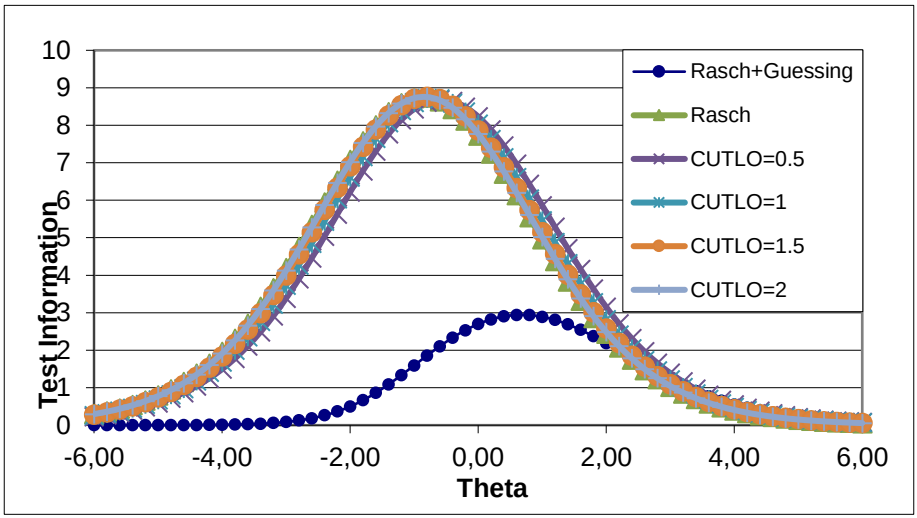

***Figure 9*. Test Information Curves for the real data.**

The test characteristic curves as presented in Figure 10 generally show that given the same ability, the expected scores for the Rasch model with guessing were higher than those from other procedures for the lower ability levels. This implies that for the same expected score, the ability estimate from the Rasch model for guessing would be lower than those from other procedures. This is consistent with the findings from a similar simulation condition with a sample size of 500 and guessing of 0.1.

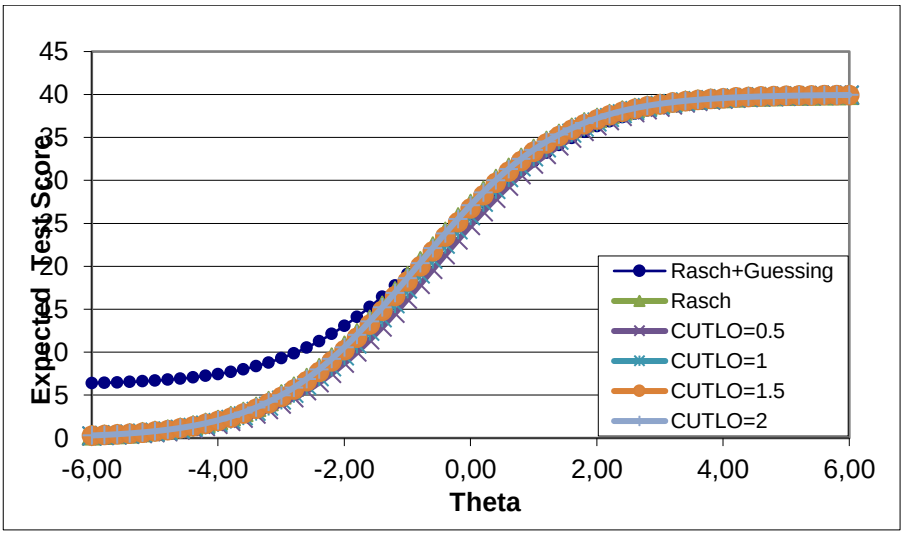

***Figure 10*. Test Characteristic Curve for the real data.**

Summary and Discussions

This study explored the model parameter estimation for the extended Rasch model for guessing using a Bayesian approach. The MCMC estimation algorithm developed in WinBUGS could well recover the true model parameters. Further, model parameter estimates were compared with those from the Rasch model which does not explicitly model the guessing effect and the Rasch model with the CUTLOW procedure to correct the guessing effects under different study conditions. The model parameter estimates were in general not significantly different between the Rasch model with and without CUTLOW correction procedures, which is consistent with the findings from Smith (2008). However, the differences in model parameter estimates from the Rasch model for guessing and other models were not negligible. The study results indicated ignoring guessing effects in general leads to the underestimation of item difficulty, overestimation of test information, and misrepresentation of the maximum test information location. The overestimation of test information may lead to premature

termination of a test in computerized adaptive test and lead to model parameter estimation errors not the same as expected.

This study explored one approach to modeling the guessing or pseudo-guessing effects in the Rasch model which is consistent with the conceptualization and parameterization of the effects in the standard IRT modeling framework such as the 3PL and 4PL IRT models. However, this model no longer maintains the specific objectivity property (Artner, 2016) and the CML estimators are no longer available. This approach incorporates a model parameter to describe the lower asymptote in the item characteristic curve. It would be interesting to compare the currently explored model with another Rasch model that specifically incorporates guessing related to the number of distractors: the multiplicative Rasch model (Smith & Fujimoto, 2011). It is worthy of note that several R-packages such as TAM, itm, mirt, and sirt are available for estimating the guessing parameters by setting some constraints using non-Bayesian estimation methods.

This study implemented the CUTLOW procedure as implemented in WINSTEPS by recoding item responses. The correction to the guessing is based on the relative difference between an examinee's ability and an item difficulty. If the item difficulty is higher than an examinee's ability by over a threshold value, the examinee's response will not contribute to the estimation of the item difficulty. This correction is not in good alignment with the inclusion of a lower asymptote in an IRT model to account for the guessing effects. It is expected that this correction procedure should be more effective in correcting ability-based guessing which could be addressed in future explorations.

When guessing is present, a three-parameter IRT model show better model parameter estimates (DeMars, 2001; Divgi, 1984). The extended Rasch model for guessing is also a potential option to model the effect. However, it is expected that in well-developed multiple-choice or non-multiple choice tests, guessing or pseudo-guessing would be limited as it is ultimately construct irrelevant variance (Smith, 2008) and not a desired item performance behavior. Good item development and test form construction is the best solution to improving item calibration quality (Gershon, 1992). On the other hand, the extended Rasch model for guessing is a convenient measurement model for analyzing item response data where guessing or pseudo-guessing factors lead to spuriously high item scores. The findings from this current exploration have more significant implications to researchers who use the Rasch model for item response data analysis when guessing is potentially present.

Guessing is a common responding behavior in item response modeling. The Rasch model is a very widely used model in test development. This model assumes no aberrant responding behaviors such as guessing be present. Though WINSTEPS, the mainstream software program for the Rasch model parameter estimation has incorporated the CUTLOW procedure to deal with the issue, more attention needs to be drawn to the impact of ignoring the guessing effects in Rasch modeling. This study compared a model extended based on the Rasch model to account for the guessing effects and compare the estimates from the model and the Rasch model with and without the

correction of the guessing effects. The findings from the study provide the measurement field with empirical evidence of the impact of the guessing effect in the Rasch modeling.

References

Andersen, E. B. (1973). Conditional inference and models for measuring. Copenhagen: Mentalhygiejnisk Forlag.

Artner, R. (2016). A simulation study of person-fit in the Rasch model. Psychological Test and Assessment Modeling, 58(3), 531-563.

Barnes, L. L. B., & Wise, S. L. (1991). The utility of a modified one-parameter IRT model with small samples. Applied Measurement in Education, 4, 143-157.

Barton, M. A. & Lord, F. M. (1981). An upper asymptote for the three-parameter logistic item-response model (Research report RR-81-20). Princeton, NJ: Educational Testing Service.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord & M. R. Novick (Eds.), Statistical theories of mental test scores (pp. 395–479). Reading, MA: Addison-Wesley.

Brooks, S. P., & Gelman, A. (1998). Alternative methods for monitoring convergence of iterative simulations. Journal of Computational and Graphical Statistics. 7, 434-455.

Cao, J., & Stokes, S. L. (2008). Bayesian IRT guessing models for partial guessing behaviors. Psychometrika, 73, 209-230.

Chen, Y. F., & Jiao, H. (2012). Detection of aberrant item respondents based on the mixture Rasch model. Paper presented at the 18th International Objective Measurement Workshop.

Choppin, B. (1983). A two-parameter latent trait model. (CSE Report No. 197). Los Angeles, CA: University of. California, Center for the Study of Evaluation.

Colonius, H. (1977). On Keats' generalization of the Rasch model. Psychometrika, 42, 443-445.

DeMars, C. (2001). Group Differences based on IRT Scores: Does the Model Matter? *Educational and Psychological Measurement, 61*, 60-70.

Dinero, T. E., & Haertel, E. (1977). Applicability of the Rasch model with varying item discriminations. Applied Psychological Measurement, 1, 581-592.

Divgi, D. R. (1984). Does small N justify use of the Rasch model? Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Fischer, G. H. (1974). Einführung in die Theorie psychologischer Tests [Introduction into the theory of psychological tests]. Bern: Huber.

Gershon, R. (1992). Guessing and measurement. Rasch Measurement Transaction, 6, 209-210.

Hessen, D. J. (2004). A new class of parametric IRT models for dichotomous item scores. Journal of Applied Measurement, 5, 385-397.

Hessen, D.J. (2005). Constant latent odds-ratios models and the Mantel-Haenszel null hypothesis. Psychometrika, 70, 497-516.

Keats, J. A. (1974). Applications of projective transformations to test theory. Psychometrika, 39,359–360.

Kubinger, K. (2005). Psychological test calibration using the Rasch model—some critical suggestions on traditional approaches. International Journal of Testing, 5, 377–394.

Kubinger, K. D., & Draxler, C. (2007). A comparison of the Rasch model and constrained Item Response Theory models for pertinent psychological test data. In M. von Davier & C. H. Carstensen (Eds.). Multivariate and mixture distribution Rasch models: Extensions and applications (pp. 293-309). New York: Springer.

Linacre, J. M. (2000). Item discrimination and infit mean-squares. Rasch Measurement Transactions, 14, 743.

Linacre, J. M. (2002). Dichotomous quasi-Rasch model with guessing. Rasch Measurement Transactions, 15, 856.

Linacre, J. M. (2004). Discrimination, guessing and carelessness: Estimating IRT parameters with Rasch. Rasch Measurement Transactions, 18, 959-960.

Linacre, J. M. (2008). Winsteps: Rasch measurement program. Winsteps.com

Lord, F. M. (1983). Maximum likelihood estimation of item response parameters when some responses are omitted. Psychometrika, 48, 477-482.

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS -- a Bayesian modeling framework: concepts, structure, and extensibility. Statistics and Computing, 10, 325--337.

McDonald, R. P. (1967). Non-linear factor analysis. Psychometric Monographs, 15.

McDonald, R. P. (1989). Future directions for item response theory. International Journal of Educational Research, 13, 205-220.

Nedelsky, L. (1954). Absolute grading standards for objective tests. Educational and Psychological Measurement, 16, 159-176.

Pelton, T. W. (2002). The accuracy of unidimensional measurement models in the presence of deviations from the underlying assumptions. Unpublished Doctoral dissertation. Brigham Young University.

Puchhammer, M. (1989). A Rasch model with guessing parameter. In K. D. Kubinger (Ed.). Modern psychometrics-A brief survey with recent contributions (pp. 271-280). Munchen: Psychologie Verlags Union.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen, Denmark: Danmarks Paedogogische Institut.

San Martin, E., Del Pino, G., & De Boeck, P. (2006). IRT models for ability-based guessing. Applied Psychological Measurement, 30, 183-203. 7.

Scheiblechner, H. H. (2009). Rasch and pseudo-Rasch models: suitableness for practical test applications. Psychology Science Quarterly, 51, 181-194.

Smith, E. V. (2008). Estimation of item difficulty for the NCLEX-RN and NCLEX-PN pretest items for inclusion in the CAT item banks. Research Report to NCSBN.

Smith, Jr., E.V., & Fujimoto, K. (2011). MultRasch: SAS code for the estimation of the Multiplicative Rasch Model parameters. Applied Psychological Measurement, 35, 485-486.

Smith, R. M. (1993). Guessing and the Rasch model. Rasch Measurement Transactions, 6, 262-263.

Thissen, D., & Steinberg, L. (1984). A response model for multiple choice items. Psychometrika, 49, 501-519.

Van de Vijver, F. J. R. (1986). The robustness of Rasch estimates. Applied Psychological Measurement, 10, 45-57.

Wainer, H., & Wright, B. (1980). Robust estimation of ability in the Rasch model. Psychometrika, 45, 373–391.

Waller, M. I. (1973). Removing the effects of random guessing from latent ability estimates. Unpublished doctoral dissertation, University of Chicago, Illinois.

Waller, M. I. (1989). Modeling guessing behavior: A comparison of two IRT models. Applied Psychological Measurement, 13, 233-243.

Weitzman, R. A. (1996). The Rasch model plus guessing. Educational and Psychological Measurement, 56, 779-790.

White, P. O. (1976). A note on Keats' generalization of the Rasch model. Psychometrika, 41, 405-407.

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: problems and potential solutions. Educational Assessment, 10(1), 1–17.

Wise, S. L., & DeMars, C. E. (2006). An application of item response time: the effort-moderated IRT model. Journal of Educational Measurement, 43(1), 19–38.

Yamamoto, K. (1995). Estimating the effects of test length and test time on parameter estimation using the HYBRID model. Research report TR-95-2. Princeton, NJ: Educational Testing Service.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). BILOG-MG 3 manual. Lincolnwood: Scientific Software International.