ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2020; 11(2); 131-146

An Evaluation of 4PL IRT and DINA Models for Estimating Pseudo-Guessing and Slipping Parameters *

Ömür Kaya KALKAN ** İsmail ÇUHADAR ***

Abstract
In an achievement test, the examinees with the required knowledge and skill on a test item are expected to answer the item correctly while the examinees with a lack of necessary information on the item are expected to give an incorrect answer. However, an examinee can give a correct answer to the multiple-choice test items through guessing or sometimes give an incorrect response to an easy item due to anxiety or carelessness. Either case may cause a bias estimation of examinee abilities and item parameters. Four-parameter logistic item response theory (4PL IRT) model and the deterministic inputs, noisy, and gate (DINA) model can be used to mitigate these negative impacts on the parameter estimations. The current simulation study aims to compare the estimated pseudo-guessing and slipping parameters from the 4PL IRT model and the DINA model under several study conditions. The DINA model was used to simulate the datasets in the study. The study results showed that the bias of the estimated slipping and guessing parameters from both 4PL IRT and DINA models were reasonably small in general although the estimated slipping and guessing parameters were more biased when datasets were analyzed through the 4PL IRT model rather than the DINA model (i.e., the average bias for both guessing and slipping parameters = .00 from DINA model, but .08 from 4PL IRT model). Accordingly, both 4PL IRT and DINA models can be considered for analyzing the datasets contaminated with guessing and slipping effects.

Key Words: 4PL IRT model, DINA model, (pseudo) guessing effect, slipping effect, lower-upper asymptote parameter.

INTRODUCTION

Psychological and educational tests are usually used for observing a sample of examinees' behaviors. Many of them focus on measuring the abilities and skills of examinees. Therefore, it is important to know how an examinee's ability determines the correctness of an answer on an item (Lord, 2012). In an achievement test, a correct response is expected from an examinee with the required knowledge on the item whereas an examinee without the necessary knowledge on the item is supposed to give an incorrect answer (Rowley & Traub, 1977). However, this assumption may not hold for the multiple-choice test items. In a test with multiple-choice test items, an examinee's response may be a reflection of true ability, guessing behavior or unexpected incorrect response (i.e., slipping effect) due to anxiety or carelessness (Liao, Ho, Yen, & Cheng, 2012; Yen, Ho, Laio, Chen, & Kuo, 2012). Under the presence of guessing and slipping effects, the estimation of examinees' abilities and item parameters might be biased. These two effects can be modeled using item response theory (IRT) models and cognitive diagnostic models (CDMs). IRT models explain the relationship between an examinee's observed test performance and its underlying latent abilities through a mathematical function (Hambleton & Swaminathan, 1985). On the other hand, CDMs are used for determining whether an examinee has a set of attributes in order to solve a problem correctly in a test (de la Torre, 2009). CDMs have many common aspects with IRT models. For example, Junker (2001), used deterministic inputs, noisy, and gate (DINA; Haertel, 1989; Junker & Sijtsma, 2001) models as an initial tool for

* We declare that a part of this study was presented as an oral presentation at the 6th International Congress on Measurement and Evaluation in Education and Psychology (CMEEP 2018) held on 5-8 September 2018 in Prizren, Kosovo.

** Assist. Prof., Pamukkale University, Faculty of Education, Denizli-Turkey, kayakalkan@pau.edu.tr, ORCID ID: 0000-0001-7088-4268
*** Ph.D., Ministry of National Education, Ankara-Turkey, ismail.cuhadar@gmail.com, ORCID ID: 0000-0002-5262-5892

To cite this article:
Kalkan, Ö. K., & Çuhadar, İ. (2020). An evaluation of 4PL IRT and DINA models for estimating pseudo-guessing and slipping parameters. *Journal of Measurement and Evaluation in Education and Psychology, 11*(2), 131-146. doi: 10.21031/epod.660273

Received: 16.12.2019
Accepted: 02.04.2020

proposing a nonparametric IRT (NIRT) for CDMs. In addition, Junker and Sijtsma (2001) showed that, as a CDM, DINA and noisy, inputs, deterministic and gate (NIDA; Maris, 1999; Junker & Sijtsma, 2001) models meet the standard assumptions of generalized multidimensional IRT models. Similarly, Meng, Xu, Zhang, and Tao (2019) showed that four-parameter logistic (4PL) (Barton & Lord, 1981) model is a special case of the higher-order DINA model with an only one latent attribute. In addition, the authors indicated that the upper asymptote in 4PL model (i.e., d_j) corresponds to the slipping parameter in CDMs (i.e., 1 - d_j). Furthermore, Culpepper (2016) stated that the lower asymptote (i.e., *c* parameter) and the upper asymptote (i.e., *d* parameter) in 4PL IRT model correspond to the guessing and slipping parameters in CDMs, respectively. Accordingly, 4PL and DINA models including (pseudo) guessing-guess and inattention-slip parameters are described shortly in the next section.

The DINA Model

DINA model, proposed by Junker and Sijtsma (2001), requires configuring a Q matrix (Tatsuoka, 1983) as the other CDM models do. This matrix is composed of ($J \times K$ times) 1 and 0s, including attributes in the columns and items in the rows of the matrix. The element in the jth row and kth column of the matrix is showed as q_{jk}. If q_{jk} equals 1, it means an examinee is required to possess the corresponding attribute in order to answer the item correctly. If the attribute is not required for answering the item correctly, q_{jk} becomes 0 in the Q matrix. Assume vector y_i represents the observed score of an examinee *i* to *J* items and the elements of y_i are statistically independent of the required attributes vector for the test $\alpha_i = \{\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{iK}\}$. Using Q-matrix and respondent's skills vector, DINA model produces the η_{ij} in Equation 1.

$$\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}} \tag{1}$$

In Equation 1, if an examinee possesses all necessary attributes for the correct answer on the item, $\eta_{ij} = 1$; otherwise, $\eta_{ij} = 0$. DINA model allows an examinee possessing all required attributes to miss an item (slip) or an examinee without at least one of the required attributes to answer the item correctly (guess). DINA model includes a guess (*g*) and slip (*s*) parameter for each test item. The parameter g_j is defined by $g_j = P(Y_{ij} = 1 \mid \eta_{ij} = 0)$, and the parameter s_j is defined by $s_j = P(Y_{ij} = 0 \mid \eta_{ij} = 1)$. Accordingly, the probability of correct response on item j given an examinee i with an attribute profile α_i is formulated as in Equation 2.

$$P(Y_{ij}=1|\alpha) = (1-s_j)^{\eta_{ij}} g_j^{1-\eta_{ij}} \tag{2}$$

DINA model can be implemented in computer software programs, including OxEdit (Doornik, 2018), LatentGold (Vermunt & Magidson, 2016), Mplus (Muthén & Muthén, 1998-2017), "CDM" package (Robitzsch, Kiefer, George, & Uenlue, 2019) and "GDINA" package (Ma & de la Torre, 2020) available as R program (R Core Team, 2017). However, it is essential to emphasize that the implementation of the DINA model is not limited to these computer software programs.

The 4PL IRT Model

Barton and Lord (1981) proposed 4PL IRT to model a parameter for the upper asymptote in the item characteristic curve. This model accounts for unexpected incorrect responses (missing) of examinees with a high ability level due to anxiety and carelessness. In the general form of this model, the probability of correct response given the ability level is formulated as in Equation 3.

$$P[X_{ij} = 1 \mid \Theta = (\theta_1, \ldots, \theta_k), a_j, b_j, c_j, d_j] = c_j + (d_j - c_j) \frac{e^{(a_{j1}\theta_1 + \ldots + a_{jk}\theta_k) - b_j}}{1 + e^{(a_{j1}\theta_1 + \ldots + a_{jk}\theta_k) - b_j}} \tag{3}$$

In Equation 3, X_{ij} is the observed score of an examinee *i* on item *j*, *k* is the number of latent factors, Θ is the vector of examinee abilities, c_j is the pseudo-guessing parameter of item *j*, d_j is the upper asymptote parameter (i.e., slipping parameter) of item *j*, a_{jk} is the discrimination parameter of item *j*

on the latent factor k, and b_j is the intercept of item j, which is the multiplication of item discrimination and item difficulty (see Barton & Lord, 1981; de Ayala, 2009). Although Barton and Lord (1981) proposed using a common upper asymptote across all test items, the general form of the 4PL model allows estimating a different upper asymptote for each test item. One-, Two-, and Three-Parameter Logistic (1PL, 2PL, and 3PL) IRT models for dichotomous items have attracted great attention in the last decade (Magis, 2013). On the other hand, 4PL IRT model was not a commonly used IRT model among practitioners and researchers until recent years due to no indication for the benefit of using 4PL IRT model, the difficulties with the estimation of upper asymptote, and the unavailability of computer software programs that can be accessed by practitioners and researchers for using 4PL IRT model (Barton & Lord, 1981; Hambleton & Swaminathan, 1985; Loken & Rulison, 2010). However, the 4PL IRT model has become more popular in recent years, especially in the literature on IRT and computerized adaptive testing (CAT), with the development of very powerful computer software programs such as the "mirt" package in R program (Chalmers, 2012; Magis, 2013; Meng et al., 2019). Many studies have contributed to the improvement of the 4PL IRT model regarding its application in the field and parameter estimation (e.g., Culpepper, 2016; Liao et al., 2012; Loken & Rulison, 2010; Magis, 2013; Meng et al., 2019; Rulison & Loken, 2009; Yen et al., 2012).

Although the conventional IRT models allow test-takers' abilities to be scaled and ordered in one or more continuous latent factors, these IRT models including 4PL IRT model are not useful to assess test-takers' strengths and weaknesses in the latent factors because IRT models do not tell if some behaviors related to the latent factors (attributes) are mastered. Unlike IRT models, CDMs were basically proposed with the purpose of identifying test-takers' strengths and weaknesses through assessing the presence or absence of several necessary attributes to solve the problems in a test (de la Torre, Hong, & Deng, 2010; de la Torre & Lee, 2010). Among CDMs, the DINA model (Junker & Sijtsma, 2001) is a commonly used model in practice and research (DeCarlo, 2011; de la Torre, 2008). Its simple and easily interpretable formula provides a good model-data fit (de la Torre & Douglas, 2008; de la Torre & Lee, 2010). Both the 4PL IRT model and the DINA model allow *c-g* and *d-s* parameters for modeling the guessing and slipping effects, respectively.

Although the literature has many studies investigating the important factors for the estimation of item parameters accurately in IRT models and CDMs separately, there are only a few studies directly comparing the item parameters from IRT models and CDMs in the same research (e.g., 2PL vs. pG-DINA in Yakar, 2017). In addition, there are some studies employing the 4PL IRT model within the CAT (e.g., Liao et al., 2012; Yen et al., 2012). However, it is also important to investigate the parameter recovery in the 4PL IRT model for a fixed (non-adaptive) test via a simulation study because the fixed tests are commonly used in educational and psychological assessments. When the similarity between IRT models and DINA model, a restricted latent model, is taken into consideration (Culpepper, 2016; Hoijtink & Molenaar, 1997; Junker, 2001; Junker & Sijtsma, 2001; Meng et al., 2019), the current study may be helpful for the field to show the similarities and differences between 4PL IRT model and DINA model, and the important study design factors for the accurate estimation of the guessing and slipping parameters. Accordingly, the current simulation study aims to compare the estimated *c-g* and *d-s* parameters from the 4PL IRT model and the DINA model using the simulated datasets through the DINA model under several study conditions.

METHOD

Simulation Study Design

All data were generated and analyzed in the R program (R Core Team, 2017). DINA model was used for data generation. In the literature, the test length was usually between 20 and 40 in many studies (e.g., Chiu, 2008; de la Torre, 2008, 2009, 2011; de la Torre & Douglas, 2004, 2008; de la Torre & Lee, 2010, 2013; Henson & Douglas, 2005). In the data generation, test length was fixed as $J = 20$ or 40 items considering these studies in the literature. The review of the literature also showed that the

studied *g* and *s* parameters tend to be between .0 and .45 (e.g., Chiu, 2008; de la Torre & Douglas, 2004; de la Torre et al., 2010; DeMars, 2007; Henson & Douglas, 2005; Huebner & Wang, 2011). In addition, the intervals of these parameters corresponding to the low, moderate, and high levels were different across the studies. In this study, three levels of *g* and *s* parameters were manipulated in the data generation: .0 - .15 (low), .15 - .30 (moderate), and .30 - .45 (high). Then, these levels were crossed between *g* and *s* parameters in the data generation. The values of *g* and *s* parameters were equally spaced with an increment of .0075 and .00375 for the conditions with 20 and 40 items, respectively. Specifically, these values were obtained taking the ratio of intervals to test length (e.g., for the test with 20 items and the parameter values between .0 and .15, .15/20 = .0075). Then, the values of *g* and *s* parameters were fixed to $g = s = .0075$ for the first item, .015 for the second item, and .15 for the last item when test length was 20, and both *g* and *s* parameters were low (.0 - .15) in the data generation. Different values were chosen for the level of correlation among factors/attributes corresponding to the weak, moderate, and strong correlations across different studies in the literature. In this study, the correlation among the attributes was fixed to $r = .2$ (weak), .5 (moderate) or .8 (strong) considering the studies by Finch (2010), and Finch, Habing, and Huynh (2003). The chosen sample size was 500, 1000, or 2000 in some simulation studies in the literature (e.g., de la Torre 2009; de la Torre & Douglas, 2004, 2008; de la Torre & Lee, 2010, 2013). However, a sample size of 1000 is sufficiently large to use the DINA model (de la Torre et al., 2010). For the 4PL IRT model, Meng et al. (2019) used a sample size of 2000. In addition, Waller and Feuerstahler (2017) found that a minimum sample size of 1000 is necessary to obtain accurate ability estimates in the 4PL IRT model. Therefore, in this study, the sample size was fixed to $N = 3000$ considering the adequacy of the sample size for the convergence of parameters to a solution. The number of attributes is usually between 4 and 8 in the literature (e.g., Chiu, 2008; de la Torre, 2011; de la Torre & Douglas, 2004; de la Torre & Lee, 2010; Huebner & Wang, 2011). Because there were many simulation conditions included in this study and the use of a great number of attributes in a simulation study is very time consuming (de la Torre & Douglas, 2004), the number of attributes was fixed to $K = 3$ or 5. Four different Q-matrices were used in the data generation (2 test lengths x 2 different numbers of attributes). Each item was linked to one attribute in all Q-matrices (one-attribute items), and the number of items was distributed across the attributes as evenly as possible. Overall, there were a total of 108 conditions for data generation (3 *g* levels x 3 *s* levels x 3 correlation levels x 2 test lengths x 2 numbers of attributes). The number of replications for each condition was 100.

Data Analysis

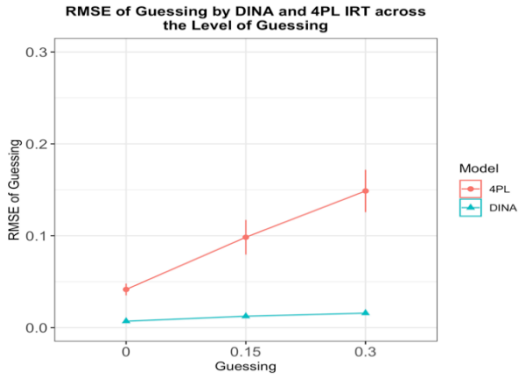
Each dataset was analyzed using a multidimensional 4PL IRT model and a DINA model. Before the analysis of datasets using the multidimensional 4PL IRT model, the dimensionality of datasets was investigated via Factor 9.2 (Lorenzo-Seva & Ferrando, 2006). Parallel analysis with the tetrachoric correlation indicated that the dimensionality assumption was met for the use of the multidimensional IRT model (i.e., it was in line with the factor structure of the datasets in the data generation via DINA model). The local independence assumption was assumed to be met because it is not within the scope of this study. Expectation-maximization (EM) algorithm was used to estimate the item parameters through 4PL IRT and DINA models because it was the default estimation method in the R packages that were used for 4PL IRT and DINA models in the study. Specifically, the analysis of datasets was conducted in the "CDM" package (Robitzsch et al., 2019) for the DINA model and the "mirt" package (Chalmers, 2012) for the 4PL model available in R program. Item-parameter bias and root mean square error (RMSE) were used to evaluate 4PL IRT and DINA models in terms of the estimation of *c-g* and *d-s* parameters correctly. 4PL IRT model was assumed to have the same true slipping and guessing parameters with the DINA model in the calculation of bias and RMSE considering the relationship between the 4PL IRT model and CDMs (see Culpepper, 2016; Meng et al., 2019). The average bias and RMSE were reported with their 95% confidence intervals across the study conditions using the formula in Equation 4.

$$\bar{\varepsilon} \pm 1.96 \frac{S_\varepsilon}{\sqrt{r_\varepsilon}} \quad (4)$$

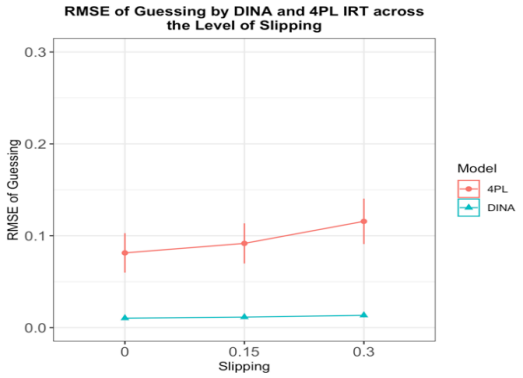
In Equation 4, $\bar{\varepsilon}$ is the average bias/RMSE of the item parameters, S_{ε} is the standard deviation of bias/RMSE of the item parameters, and r_{ε} is the number of study conditions when calculating the average bias/RMSE of the item parameters.

RESULTS

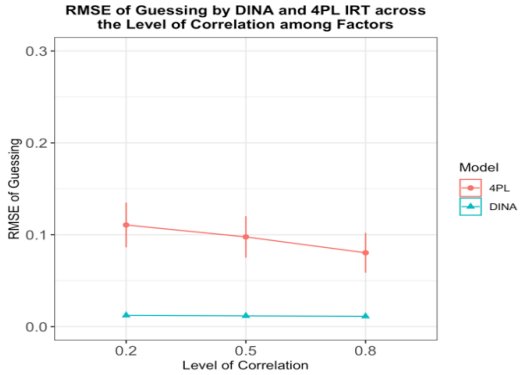
Results were summarized using the average RMSE of the item parameters and creating its 95% confidence intervals by the 4PL IRT and DINA models across the study conditions. The RMSE of guessing parameters are presented across 4PL and DINA models in Figure 1. The RMSE of the guessing parameters were almost zero across all levels of *c*-*g* parameters (*c*-*g* parameters = .0, .15, and .3; see Figure 1a), all levels of *d*-*s* parameters (*d*-*s* parameters = .0, .15, and .3; see Figure 1b), all levels of the correlation among factors/attributes (r = .2, .5, and .8; see Figure 1c), all numbers of attributes (K = 3 and 5; see Figure 1d), and all test lengths (J = 20 and 40; see Figure 1e) in the study when DINA model was fit to the data.



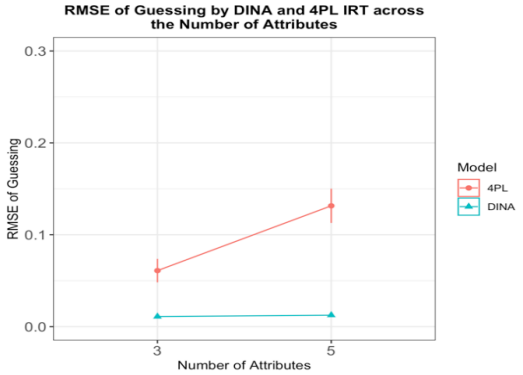
(a)



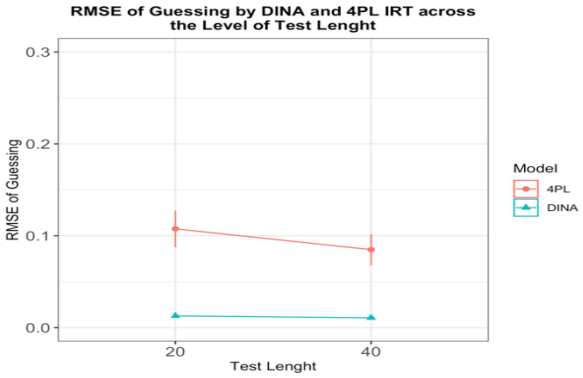
(b)



(c)



(d)

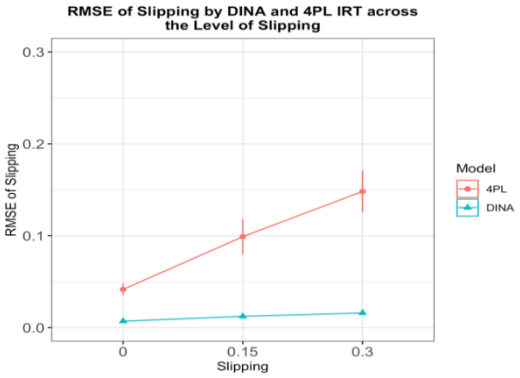


(e)

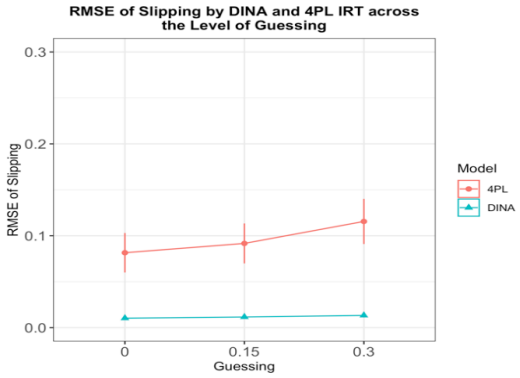
Note. On x axis of Figure 1a and 1b, 0 = the values between .0 and .15 (low); 0.15 = the values between .15 and .30 (moderate); 0.3 = the values between .30 and .45 (high).

Figure 1. The 95% Confidence Intervals of (Pseudo) Guessing-parameter RMSE by DINA and 4PL IRT Models across Different Study Conditions

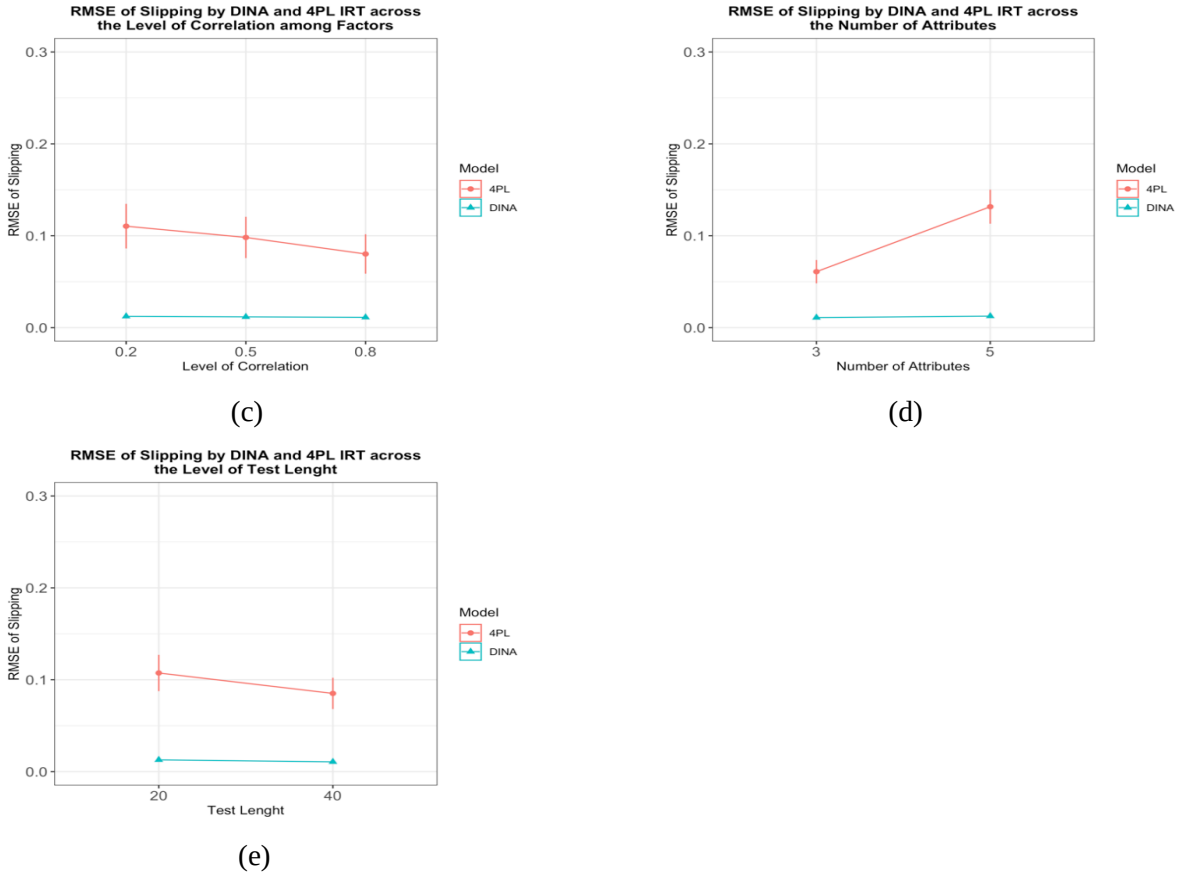
In addition, its 95% confidence intervals were so small across all these study conditions that they did not appear in any figure for DINA models. However, the average RMSE of the guessing parameters became larger across all study conditions when the 4PL IRT model was fit to the data in lieu of the DINA model (see Figure 1a, 1b, 1c, 1d, and 1e). Furthermore, the RMSE of the guessing parameters were larger for 4PL IRT model under the conditions with a larger *c-g* parameter in the data generation (the 95% confidence interval of the average RMSE for the guessing parameters was between .04 and .05 when *c-g* parameters = .0, between .08 and .12 when *c-g* parameters = .15, and between .13 and .17 when *c-g* parameters = .3; see Figure 1a). Similarly, for 4PL IRT model, the average RMSE of the guessing parameters became larger when the number of factors/attributes was greater, the test was shorter, *d-s* parameters were higher, and the correlation among factors/attributes was weaker, as expected (see Figure 1b, 1c, 1d, and 1e). However, among these four study conditions, the number of factors/attributes was the only significant study condition for the size of the RMSE of the guessing parameters from 4PL IRT model when the overlap between the 95% confidence intervals was considered (the 95% confidence interval of the average RMSE for the guessing parameters was between .05 and .07 when K = 3, and between .11 and .15 when K = 5; see Figure 1d). Overall, the similar results were also found for the RMSE of the slipping parameters (see Figure 2).



(a)



(b)

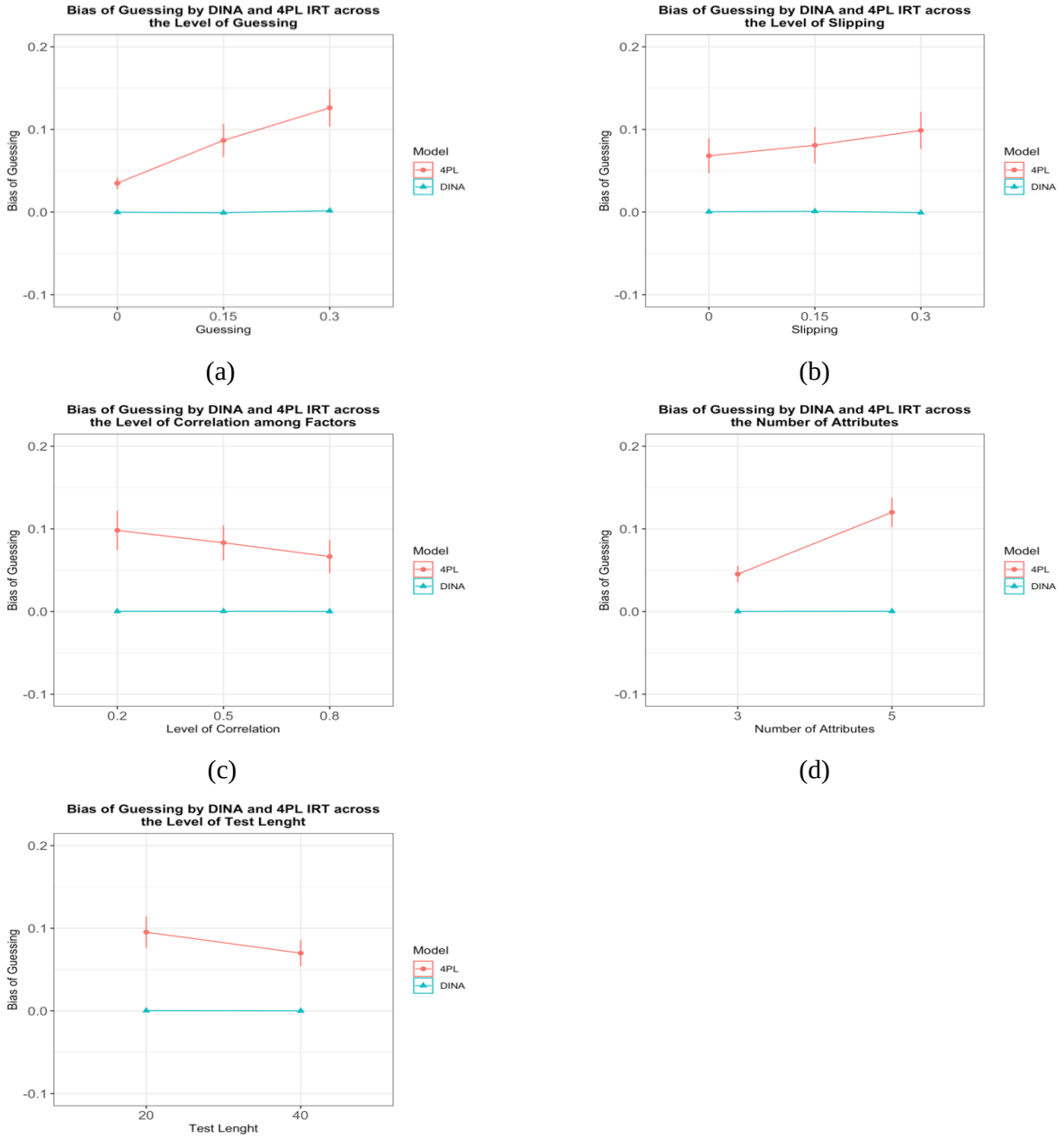


Note. On x axis of Figure 2a and 2b, 0 = the values between .0 and .15 (low); 0 .15 = the values between .15 and .30 (moderate); 0.3 = the values between .30 and .45 (high).

Figure 2. The 95% Confidence Intervals of Slipping-parameter RMSE by DINA and 4PL IRT Models across Different Study Conditions

The average RMSE of the slipping parameters with its confidence interval was almost identical to the RMSE of the guessing parameters across all study conditions for both DINA and 4PL IRT models with one exception (see Figure 2b, 2c, 2d, and 2e). The RMSE of the slipping parameters became larger for 4PL IRT model under the conditions with a larger *d-s* parameter rather than *c-g* parameter in the data generation considering the confidence intervals of average RMSEs across the study conditions (the 95% confidence interval of the average RMSE for the slipping parameters was between .04 and .05 when *d-s* parameters = .0, between .08 and .12 when *d-s* parameters = .15, and between .13 and .17 when *d-s* parameters = .3; see Figure 2a).

The bias of the guessing and slipping parameters were calculated as the expectation of the difference between the item parameters estimated from DINA or 4PL IRT models and their corresponding values from the true model in the data generation. Results were summarized using the average bias of the item parameters and creating its 95% confidence intervals by 4PL IRT and DINA models across the study conditions. The bias of guessing parameters are presented across 4PL and DINA models in Figure 3.

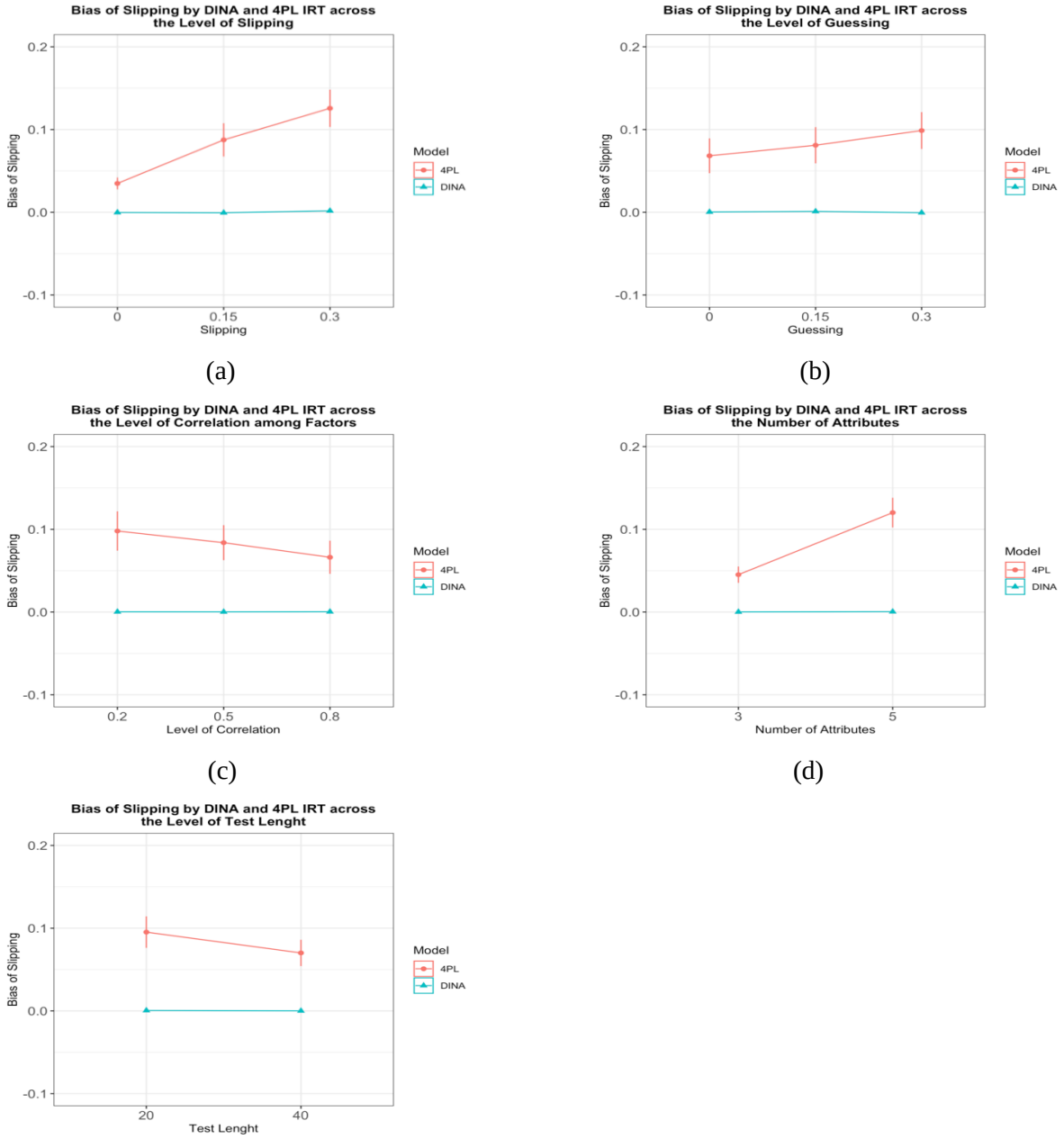


Note. On x axis of Figure 3a and 3b, 0 = the values between .0 and .15 (low); 0.15 = the values between .15 and .30 (moderate); 0.3 = the values between .30 and .45 (high).

Figure 3. 95% Confidence Intervals of (Pseudo) Guessing-parameter Bias by DINA and 4PL IRT Models across Different Study Conditions

As expected from the RMSEs of the guessing parameters, when the guessing parameters were estimated through DINA model, the bias of the guessing parameters were almost zero with a very narrow confidence interval across all levels of *c-g* parameters (*c-g* = .0, .15, and .3; see Figure 3a), all levels of *d-s* parameters (*d-s* = .0, .15, and .3; see Figure 3b), all levels of the correlation among factors/attributes (*r* = .2, .5, and .8; see Figure 3c), all numbers of attributes (*K* = 3 and 5; see Figure 3d), and all test lengths (*J* = 20 and 40; see Figure 3e) in the study. Unlike the DINA model, the guessing parameters were overestimated across all study conditions when the 4PL IRT model was used to estimate the guessing parameters (see Figure 3a, 3b, 3c, 3d, and 3e). In addition, the overestimation of the guessing parameters became more severe for the 4PL IRT model under the

conditions with a higher *c-g* parameter, a higher *d-s* parameter, a weaker correlation among factors/attributes, a greater number of factors/attributes, and a shorter test in the data generation.



Note. On x axis of Figure 4a and 4b, 0 = the values between .0 and .15 (low); 0.15 = the values between .15 and .30 (moderate); 0.3 = the values between .30 and .45 (high).

Figure 4. 95% Confidence Intervals of Slipping-parameter Bias by DINA and 4PL IRT Models across Different Study Conditions

However, among these study conditions, the value of *c-g* parameter and the number of factors/attributes in the data generation were the only study conditions that made a significant difference on the bias of the guessing parameters from 4PL IRT model considering the overlap between the 95% confidence intervals (the 95% confidence interval of the average bias for guessing

parameters was between .03 and .04 when *c-g* parameters = .0, between .07 and .11 when *c-g* parameters = .15, and between .10 and .15 when *c-g* parameters = .3; between .04 and .05 when K = 3, and between .10 and .14 when K = 5; see Figure 3a and Figure 3d, respectively). The similar results were also found for the bias of the slipping parameters (see Figure 4). However, like the RMSE of the slipping parameters, the overestimation of the slipping parameters were more severe under the conditions with a larger *d-s* parameter rather than a larger *c-g* parameter in the data generation when the 95% confidence intervals of the average bias for the slipping parameters were taken into consideration across the study conditions (i.e., the 95% confidence interval of the average bias for slipping parameters was between .03 and .04 when *d-s* parameters = .0, between .07 and .11 when *d-s* parameters = .15, and between .10 and .15 when *d-s* parameters = .3; but the 95% confidence interval of the average bias for slipping parameters was between .05 and .09 when *c-g* parameters = .0, between .06 and .10 when *c-g* parameters = .15, and between .08 and .12 when *c-g* parameters = .3; see Figure 4a and 4b).

DISCUSSION and CONCLUSION

Multiple-choice test items might be regarded as a popular item type in educational and psychological assessments. However, in a test with multiple-choice test items, some test takers may guess a correct answer (i.e., guessing effect), or miss it because of anxiety or carelessness (i.e., slipping effect). The estimation of item parameters and test-takers' abilities might be biased when the guessing effect and/or the slipping effect are not modeled in data analyses. The DINA model and 4PL IRT model consider the guessing and slipping effects through including a parameter for the guessing effect (i.e., *g* parameter in DINA model and *c* parameter in 4PL IRT model) and a parameter for the slipping effect (i.e., *s* parameter in DINA model and *d* parameter in 4PL IRT model) when analyzing data and estimating model parameters such as item parameters and test-takers' abilities. The current simulation study purported to compare the estimated *c-g* and *d-s* parameters from the 4PL IRT model and DINA model through manipulating the number of attributes, the level of correlation among attributes, test length, the level of *g* parameter, and the level of *s* parameter.

The research findings indicate that the guessing and slipping parameters were estimated correctly across all study conditions when the DINA model was used to analyze the datasets in the study (e.g., the RMSEs of the guessing and slipping parameters were almost zero across all study conditions). The good performance of the DINA model is consistent with the results in the literature (e.g., Chiu, 2008; de la Torre et al., 2010; de la Torre & Lee, 2010). However, an important limitation of the current study is the use of the DINA model for data generation. Fitting the correct model (i.e., DINA model) might be a possible reason for the estimation of slipping and guessing parameters correctly. Thus, it might be helpful to use an empirical dataset for the evaluation of guessing and slipping parameters estimated via 4PL IRT and DINA models in a future study.

A typical test length is 15 or 20 to estimate the model parameters accurately in the CDMs, and the model parameters are estimated more accurately via the DINA model as the sample size becomes larger (de la Torre, 2009; de la Torre et al., 2010). In the current study, the test length was fixed as 20 or 40 items, and the sample size was fixed at 3000 in the data generation. The large sample size and the long test length might be other possible reasons for the estimation of slipping and guessing parameters accurately via the DINA model. Future work may consider investigating the impact of a shorter test length (e.g., < 15 or 20) and a smaller sample size (e.g., < 3000) on the accuracy of guessing and slipping parameters estimated via 4PL IRT and DINA models.

Both guessing and slipping parameters were overestimated when the 4PL IRT model was chosen to estimate these two item parameters in lieu of the DINA model. The number of attributes made a significant difference in the overestimation of both guessing and slipping parameters when the 4PL IRT model was fit to the data. The overestimation of the guessing and slipping parameters from the 4PL IRT model became more severe when the number of attributes was greater in the data generation. While the number of attributes became greater for the conditions with the same test length, there were fewer items per attribute. Parameter estimates tend to be more biased for a shorter test (Hulin, Lissak,

& Drasgow, 1982). This might be a possible reason for the overestimation of the guessing and slipping parameters more severely under the conditions with a greater number of attributes given the same test length.

The value of guessing parameters in the data generation was another significant study condition for the estimation of guessing parameters through the 4PL IRT model. The guessing parameters were overestimated more under the conditions with a larger guessing parameter in the data generation. This was not consistent with the results from DeMars' (2007) study where the overestimation was more severe for the conditions with a lower guessing parameter. DeMars fits a unidimensional 3PL IRT model to the datasets that followed a multidimensional 3PL IRT model whereas we analyzed the datasets with the multidimensional factor structure and the slipping effect through fitting a multidimensional 4PL IRT model to the datasets. In addition, due to the small sample size (i.e., 1000), the estimated guessing parameters were biased towards the mean of prior distribution (i.e., .2) in DeMars' study (i.e., the bias = .05, .02, .01, -.01, and -.03 for c = .10, .15, .20, .25, and .30, respectively). However, a relatively larger sample size (i.e., 3000) was used in the current study. These might be some possible reasons for the difference between the findings. Although the average bias of the guessing parameters became larger for the 4PL IRT model under the conditions with a higher slipping parameter, a weaker correlation among attributes, and a shorter test in the data generation, the bias difference was not significant considering the overlap between the 95% confidence intervals. This is consistent with the findings in the literature considering the impact of test length and correlation among attributes (e.g., Hulin et al., 1982; Svetina, Valdivia, Underhill, Dai, & Wang, 2017).

When the slipping parameters were estimated through the 4PL IRT model, the overestimation of slipping parameters was more severe under the conditions with a greater slipping parameter in the data generation. However, the bias of the slipping parameters from the 4PL IRT model did not differ across the different levels of the guessing parameters, the correlation among attributes, and the test length in the data generation when the 95% confidence interval of the average bias was taken into consideration. The findings related to the estimated slipping parameters may not be generalized to other study conditions, and there is a need for more studies investigating the parameter recovery in the 4PL IRT model under different study conditions. For example, as mentioned before, the sample size was not manipulated in the current study, and the chosen sample size was limited to 3000 for data generation. However, it is common to use a sample size less than 3000 in literature (see Conway & Huffcutt, 2003; Henson & Roberts, 2006; Jackson, Gillaspy, & Purc-Stephenson, 2009). Although it is recommended that the sample size for running a 3PL model or a DINA model should be larger than 1000 to obtain accurate parameter estimates, there is no rule of thumb for the required sample size of the 4PL IRT model (de la Torre et al., 2010; Hulin et al., 1982). Accordingly, the sample size (e.g., < 3000) might be manipulated in future work to investigate the lower limit for the sample size for running a 4PL IRT model. In addition, it might be helpful to study whether the manipulation of sample size will make a difference in the estimation of slipping and guessing parameters by interacting with the other study conditions such as test length and the correlation among attributes.

Although the estimated slipping and guessing parameters were more biased when datasets were analyzed through the 4PL IRT model than the DINA model, the bias of the estimated slipping and guessing parameters from both 4PL IRT and DINA models were reasonably small in general. Overall, the average bias of both guessing and slipping parameters was smaller than .1 across all study conditions, except the conditions with a high guessing/slipping parameter or a great number of attributes in the data generation. Accordingly, both 4PL IRT and DINA models can be preferred for analyzing the datasets contaminated with guessing and slipping effects. However, it is important to consider the aforementioned limitations of the current simulation study before deciding whether the study results can be generalized to other study settings.

Compliance with Ethical Standards

Funding

This research was supported by Pamukkale University Scientific Research Projects Coordination Unit under code ADEP-2018KRM002-063.

REFERENCES

Barton, M. A., & Lord, F. M. (1981). *An upper asymptote for the three-parameter logistic item-response model* (Research Report 18-21). Princeton, NJ: Educational Testing Service. doi: 10.1002/j.2333-8504.1981.tb01255.x

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1-29. doi: 10.18637/jss.v048.i06

Chiu, C. Y. (2008). *Cluster analysis for cognitive diagnosis: Theory and applications* (Doctoral dissertation). Retrieved from https://www.ideals.illinois.edu/handle/2142/80055

Conway, J. M., & Huffcutt, A. I. (2003). A review and evaluation of exploratory factor analysis practices in organizational research. *Organizational Research Methods*, *6*(2), 147-168. doi: 10.1177/1094428103251541

Culpepper, S. A. (2016). Revisiting the 4-parameter item response model: Bayesian estimation and application. *Psychometrika*, *81*(4), 1142-1163. doi: 10.1007/s11336-015-9477-6

de Ayala, R. J. (2009). *The theory and practice of item response theory.* New York, NY: The Guilford Press.

DeCarlo, L. T. (2011). On the analysis of fraction subtraction data: The DINA model, classification, latent class sizes, and the Q-matrix. *Applied Psychological Measurement*, *35*(1), 8-26. doi: 10.1177/0146621610377081

de la Torre, J. (2008). An empirically based method of Q-matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement*, *45*(4), 343-362. doi: 10.1111/j.1745-3984.2008.00069.x

de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational And Behavioral Statistics*, *34*(1), 115-130. doi: 10.3102/1076998607309474

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*(2), 179-199. doi: 10.1007/s11336-011-9207-7

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrica*, *69*(3), 333-353. doi: 10.1007/BF02295640

de la Torre, J., & Douglas, J. A. (2008). Model evaluation and multiple strategies in cognitive diagnosis: An analysis of fraction subtraction data. *Psychometrika*, *73*(4), 595-624. doi: 10.1007/s11336-008-9063-2

de la Torre, J., Hong, Y., & Deng, W. (2010). Factors affecting the item parameter estimation and classification accuracy of the DINA model. *Journal of Educational Measurement*, *47*(2), 227-249. doi: 10.1111/j.1745-3984.2010.00110.x

de la Torre, J., & Lee, Y. S. (2010). A note on the invariance of the dina model parameters. *Journal of Educational Measurement*, *47*(1), 115-127. doi: 10.1111/j.1745-3984.2009.00102.x

de la Torre, J., & Lee, Y. S. (2013). Evaluating the Wald test for item- level comparison of saturated and reduced models in cognitive diagnosis. *Journal of Educational Measurement*, *50*(4), 355-373. doi: 10.1111/jedm.12022

DeMars, C. E. (2007). "Guessing" parameter estimates for multidimensional item response theory models. *Educational and Psychological Measurement, 67*(3), 433-446. doi: 10.1177/0013164406294778

Doornik, J. A. (2018). *An object-oriented matrix programming language Ox (Version 8.0)* [Computer software]. London: Timberlake Consultants Press.

Finch, H. (2010). Item parameter estimation for the MIRT model: Bias and precision of confirmatory factor analysis-based models. *Applied Psychological Measurement*, *34*(1), 10-26. doi: 10.1177/0146621609336112

Finch, H., Habing, B. T., & Huynh, H. (2003, April). *Comparison of NOHARM and conditional covariance methods of dimensionality assessment*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26*(4), 301-321. doi: 10.1111/j.1745-3984.1989.tb00336.x

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer.

Henson, R., & Douglas, J. (2005). Test construction for cognitive diagnosis. *Applied Psychological Measurement*, *29*(4), 262-277. doi: 10.1177/0146621604272623

Henson, R. K., & Roberts, J. K. (2006). Use of exploratory factor analysis in published research: Common errors and some comment on improved practice. *Educational and Psychological Measurement, 66*(3), 393-416. doi: 10.1177/0013164405282485

Hoijtink, H., & Molenaar, I. W. (1997). A multidimensional item response model: Constrained latent class analysis using the Gibbs sampler and posterior predictive checks. *Psychometrika*, *62*(2), 171-189. doi: 10.1007/BF02295273

Huebner, A., & Wang, C. (2011). A note on comparing examinee classification methods for cognitive diagnosis models. *Educational and Psychological Measurement*, *71*(2), 407-419. doi: 10.1177/0013164410388832

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two- and three-parameter logistic item characteristic curves: A monte carlo study. *Applied Psychological Measurement, 6*(3), 249-260. doi: 10.1177/014662168200600301

Jackson, D. L., Gillaspy, J. A., & Purc-Stephenson, R. (2009). Reporting practices in confirmatory factor analysis: an overview and some recommendations. *Psychological Methods, 14*(1), 6-23. doi: 10.1037/a0014694

Junker, B. W. (2001). On the interplay between nonparametric and parametric IRT, with some thoughts about the future. In A. Boomsma, M. A. J. Van Duijn, &T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 274-276). New York, NY: Springer-Verlag.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*(3), 258-272. doi: 10.1177/01466210122032064

Liao, W. W., Ho, R. G., Yen, Y. C., & Cheng, H. C. (2012). The four-parameter logistic item response theory model as a robust method of estimating ability despite aberrant responses. *Social Behavior and Personality: An International Journal*, *40*(10), 1679-1694. doi: 10.2224/sbp.2012.40.10.1679

Loken, E., & Rulison, K. L. (2010). Estimation of a four-parameter item response theory model. *British Journal of Mathematical and Statistical Psychology*, *63*(3), 509-525. doi: 10.1348/000711009X474502

Lorenzo-Seva, U., & Ferrando, P. J. (2006). FACTOR: A computer program to fit the exploratory factor analysis model. *Behavior Research Methods, Instruments, & Computers, 38*(1), 88-91. doi: 10.3758/BF03192753

Lord, F. M. (2012). *Applications of item response theory to practical testing problems*. New Jersey, NJ: Lawrence Erlbaum Associates.

Ma, W., & de la Torre, J. (2020). *GDINA: The generalized DINA model framework*: *R package (Version 2.7.9).* Retrieved from https://CRAN.R-project.org/package=GDINA

Magis, D. (2013). A note on the item information function of the four-parameter logistic model. *Applied Psychological Measurement, 37*(4), 304-315. doi: 10.1177/0146621613475471

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika, 64*, 187-212. doi: 10.1007/BF02294535

Meng, X., Xu, G., Zhang, J., & Tao, J. (2019). Marginalized maximum a posteriori estimation for the four-parameter logistic model under a mixture modelling framework. *British Journal of Mathematical and Statistical Psychology,* Advanced online publication. doi: 10.1111/bmsp.12185

Muthén, L. K., & Muthén, B. O. (1998-2017). *Mplus user's guide* (8th ed.). Los Angeles, CA: Muthén & Muthén.

R Core Team. (2017). *R: A language and environment for statistical computing* [Computer Software]. Vienna, Austria: R Foundation for Statistical Computing.

Robitzsch, A., Kiefer, T., George, A. C., & Uenlue, A. (2019). *Package 'CDM'.* Retrieved from https://cran.r-project.org/web/packages/CDM/CDM.pdf

Rowley, G. L., & Traub, R. E. (1977). Formula scoring, number-right scoring, and test-taking strategy. *Journal of Educational Measurement*, *14*(1), 15-22. doi: 10.1111/j.1745-3984.1977.tb00024.x

Rulison, K. L., & Loken, E. (2009). I've fallen and i can't get up: Can high-ability students recover from early mistakes in CAT? *Applied Psychological Measurement*, *33*(2), 83-101. doi: 10.1177/0146621608324023

Svetina, D., Valdivia, A., Underhill, S., Dai, S., & Wang, X. (2017). Parameter recovery in multidimensional item response theory models under complexity and nonnormality. *Applied Psychological Measurement, 41*(7), 530-544. doi: 10.1177/0146621617707507

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. doi: 10.1111/j.1745-3984.1983.tb00212.x

Vermunt, J. K., & Magidson, J. (2016). *Upgrade manual for latent GOLD 5.1.* Belmont, MA: Statistical Innovations Inc.

Waller, N. G., & Feuerstahler, L. (2017). Bayesian modal estimation of the four-parameter item response model in real, realistic, and idealized data sets. *Multivariate behavioral research*, *52*(3), 350-370. doi: 10.1080/00273171.2017.1292893

Yakar, L. (2017). *Bilişsel tanı ve çok boyutlu madde tepki kuramı modellerinin karşılıklı uyumlarının incelenmesi* (Doctoral thesis). Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi/

Yen, Y. C., Ho, R. G., Laio, W. W., Chen, L. J., & Kuo, C. C. (2012). An empirical evaluation of the slip correction in the four parameter logistic models with computerized adaptive testing. *Applied Psychological Measurement, 36*(2), 75-87. doi: 10.1177/0146621611432862

Tahmin ve Kaydırma Parametrelerinin Kestiriminde 4PL MTK ve DINA Modellerinin Değerlendirilmesi

Giriş

Psikolojik veya eğitimsel testler genellikle adayların bir davranış örneklemini gözlemlemek için kullanılmaktadır. Bu testlerin birçoğu adayların yetenek veya beceri düzeylerinin ölçülmesine odaklanmaktadır. Bu nedenle bir adayın yeteneğinin, bir maddenin doğru cevaplanmasını nasıl belirlediğinin bilinmesi oldukça önemlidir (Lord, 2012). Genellikle bir başarı testinde gerekli bilgiye sahip adayların maddeyi doğru cevaplamaları, sahip olmayanların ise yanlış cevaplamaları beklenir (Rowley & Traub, 1977). Ancak çoktan seçmeli testlerde bu varsayım her zaman geçerli olmayabilir. Bireyin çoktan seçmeli testlerde verdiği cevaplarda; gerçek yeteneğin yansıması görülebilir, doğru cevaba şans başarısı ile ulaşabilir ya da endişe veya dikkatsizlikten kaynaklı yanlış cevaplar görülebilir (Liao, Ho, Yen, & Cheng, 2012; Yen, Ho, Laio, Chen, & Kuo, 2012). Son iki durumda bireylerin yetenek ve madde parametre kestirimleri yanlı olabilir. Bu durum bazı madde tepki kuramı (IRT) ve bilişsel tanı modelleri (CDMs) tarafından ele alınmaktadır. Şans başarısı-tahmin (Pseudo guessing-guess, *c-g*) ve dikkatsizlik-kaydırma (inattention-slip, *d-s*) parametrelerini ele alan 4 parametreli lojistik (4PL) (Barton & Lord, 1981) model ve DINA (Haertel, 1989; Junker & Sijtsma, 2001) model, bu modellere örnek verilebilir. Bu araştırmanın amacı DINA modele uygun olarak farklı koşullarda üretilen veriler üzerinden 4PL Madde Tepki Kuramı (MTK) ve DINA modelleriyle elde edilen *c-g* ve *d-s* parametrelerini karşılaştırmaktır. Böylece her iki model arasındaki farklılıkların ve benzerliklerin ortaya konulması, *c-g* ve *d-s* doğru parametre kestirimini etkileyen faktörlerin belirlenmesi ve bu parametre tasarımlarına sahip araştırmalara katkıda bulunulması amaçlanmıştır.

Yöntem

Verilerin üretimi ve analizi R yazılımı (R Core team, 2017) ile gerçekleştirilmiştir. Veriler DINA modele uygun olarak üretilmiştir. Bu çalışmadaki koşullar belirlenirken literatürde yer alan çalışmalar dikkate alınmıştır (örn., Chiu, 2008; de la Torre, 2008, 2009, 2011; de la Torre & Douglas, 2004, 2008; de la Torre & Lee, 2010, 2013; de la Torre, Hong, & Deng, 2010; DeMars, 2007; Finch, 2010; Finch, Habing, & Huynh, 2003; Henson & Douglas, 2005; Huebner & Wang, 2011; Meng, Xu, Zhang, & Tao, 2019; Waller & Feuerstahler, 2017). Bu doğrultuda veri üretiminde J = 20 ve J = 40 test uzunlukları dikkate alınmıştır. Bunun yanı sıra .0-.15 (düşük), .15-.30 (orta) ve .30-.45 (yüksek) olmak üzere 3 farklı *g* ve *s* parametre düzeyi belirlenmiştir. Özellikler arası korelasyon düzeyleri r = .2 (düşük), r = .5 (orta), ve r = .8 (yüksek) olarak belirlenmiştir. Modellerden elde edilen parametrelerin doğruluğu için örneklem büyüklüğü N = 3000'e sabitlenmiştir. Ayrıca iki farklı özellik sayısı K = 3 ve K = 5 dikkate alınmıştır. Veri üretiminde dört farklı Q-matris kullanılmıştır (2 test uzunluğu x 2 özellik sayısı). Q-matrislerde yer alan her bir madde bir özellik ile ilişkilendirilmiştir. Q-matrislerde yer alan özellikler ile ilişkili madde sayılarının eşit olmasına dikkat edilmiştir. Araştırma kapsamında toplam 108 koşul (3 *g* düzeyi x 3 *s* düzeyi x 3 korelasyon düzeyi x 2 test uzunluğu x 2 özellik sayısı) test edilmiştir. Her bir koşul için 100 veri seti üretilmiştir. Her bir veri seti çok boyutlu 4PL MTK ve DINA modeller ile analiz edilmiştir. Çok boyutlu 4PL MTK'nın veri analizi için uygunluğunu test

etmek için verilerin faktör yapısı Factor 9.2 yazılımı (Lorenzo-Seva & Ferrando, 2006) ile incelenmiştir. Tetrakorik korelasyona dayalı paralel analizler sonucunda, çok boyutluluk varsayımının kullanılan MTK modeline uygun olduğu doğrulanmıştır. Bu çalışma kapsamı dışında olması nedeniyle üretilen verilerin yerel bağımsızlık varsayımını karşıladığı varsayılarak analizler gerçekleştirilmiştir. DINA model analizleri "CDM" (Robitzsch, Kiefer, George, & Uenlue, 2019) paketi ile gerçekleştirilmiştir. 4PL analizleri için "mirt" (Chalmers, 2012) paketi kullanılmıştır. 4PL MTK ve DINA modellerin *c*-*g* ve *d*-*s* parametre kestirimlerinin doğruluğunun değerlendirilmesinde sapma (bias) ve hata kareler ortalaması karekökü (RMSE) değerleri kullanılmıştır. Sapma ve RMSE değerleri hesaplanırken 4PL MTK'nın tahmin ve kaydırma parametrelerinin DINA modeli ile aynı gerçek değere sahip olduğu varsayılmıştır (geniş bilgi için bkz., Culpepper, 2016; Meng ve diğerleri, 2019). Ortalama sapma ve RMSE değerleri %95 güven aralıkları ile rapor edilmiştir.

Sonuç ve Tartışma

Araştırma kapsamında ulaşılan bulgular, DINA modeli kullanıldığında tahmin (şans başarısı) ve kaydırma parametrelerinin ele alınan tüm çalışma koşullarında doğru bir şekilde kestirildiğini ortaya koymuştur. Tüm çalışma koşulları altında DINA modeli kullanıldığında tahmin ve kaydırma parametrelerinin RMSE değerleri sıfıra yakın bulunmuştur. DINA modelin parametre kestiriminde iyi bir performans sergilemesi literatürdeki diğer çalışma sonuçlarıyla uyumludur (örn., Chiu, 2008; de la Torre & Lee, 2010; de la Torre ve diğerleri, 2010). Ancak, veri üretiminde DINA model kullanılması bu çalışmanın önemli bir sınırlılığıdır. Tahmin ve kaydırma parametrelerinin doğru kestirimi, veriler analiz edilirken doğru model olan DINA modelinin kullanılmasından kaynaklanmış olabilir. Bu nedenle 4PL MTK ve DINA modellerinin tahmin ve kaydırma parametrelerinin kestirimi açısından karşılaştırılması için gelecek çalışmalarda gerçek veri setinin kullanılması önerilmektedir.

CDM modellerinde parametrelerin doğru kestirimi için tipik bir test uzunluğunun 15 ila 20 olduğu ve örneklem büyüklüğü arttıkça DINA modeli kullanılarak yapılan parametre kestirimlerinin daha doğru sonuçlar verdiği bilinmektedir (de la Torre, 2009; de la Torre ve diğerleri, 2010). Bu çalışmada veri üretiminde test uzunluğu 20 ve 40 olarak belirlenmiş ve örneklem büyüklüğü 3000'de sabitlenmiştir. Örneklem büyüklüğünün ve test uzunluklarının yeterli olmasının tahmin ve şans parametrelerinin DINA model kestirim doğruluklarında etkili olduğu düşünülmektedir. Bu nedenle sonraki çalışmalarda test uzunluğunun daha kısa tutulmasının ve düşük örneklem büyüklüklerinin söz konusu sonuçlarda ne gibi değişikliklere neden olacağı incelenebilir.

DINA model yerine 4PL MTK modeli kullanıldığında hem tahmin hem de kaydırma parametresinin gerçek değerlerinden daha büyük kestirimlere neden olduğu belirlenmiştir. Bu durumda özellik sayısının önemli olduğu ve özellik sayısı arttıkça tahmin ve kaydırma parametrelerinin 4PL MTK ile kestirilen değerlerinin gerçek değerlerinden daha da uzaklaştığı bulunmuştur. Test uzunluğu sabit tutularak özellik sayısı artırıldığında her bir özellikle ilişkilendirilmiş madde sayısı azalmaktadır. Bu nedenle daha kısa testlerde parametre kestirimi daha yanlı olmaktadır (Hulin, Lissak, & Drasgow, 1982). Bu doğrultuda test sabit tutulurken özellik sayısının artırılmasının tahmin ve kaydırma parametrelerinde daha yanlı kestirimlere neden olduğu düşünülebilir.

Tahmin parametresinin veri üretimindeki değerinin büyük olması 4PL MTK modeliyle kestirilen tahmin parametresinin daha yanlı olmasına neden olmuştur. Benzer şekilde kaydırma parametresinin veri üretimindeki değerini büyütmek, 4PL MTK modeliyle kestirilen kaydırma parametresinin daha yanlı olmasıyla sonuçlanmıştır. Ancak %95 güven aralıkları dikkate alındığında söz konusu parametrelerin özellikler arası korelasyondan ve test uzunluğundan kayda değer bir şekilde etkilenmediği bulunmuştur. Bu sonuç, test uzunluğu ve özellikler arası korelasyon gibi çalışma özellikleri açısından literatürde bulunan sonuçlarla örtüşmektedir (örn., Hulin ve diğerleri, 1982; Svetina, Valdivia, Underhill, Dai, & Wang, 2017).

Her ne kadar 4PL MTK modeliyle elde edilen tahmin ve kaydırma parametreleri DINA modele kıyasla daha yanlı olsa da, bu kestirimlerdeki yanlılığın genel anlamda önemli olmadığı söylenebilir. Örneğin, tüm çalışma koşulları dikkate alındığında tahmin ve kaydırma parametrelerindeki ortalama yanlılığın

genel olarak .1'den küçük olduğu bulunmuştur. Sadece tahmin ve kaydırma parametrelerinin veri üretimindeki değerlerinin yüksek olduğu koşullar ile özellik sayısının büyük olduğu çalışma koşullarında 4PL MTK modeliyle yapılan kestirimlerin yanlılığı .1'den büyük bulunmuştur. Bu sonuçlar dikkate alındığında araştırmacılar tahmin ve kaydırma etkisine sahip verilerin analizlerinde hem DINA modelini hem de 4PL MTK modelini dikkate alabilirler. Ancak bu sonuçları başka çalışma koşullarına genellemeden önce çalışmanın sınırlılıklarının dikkate alınması oldukça önemlidir.

Yukarıda bahsedilen çalışma sınırlıkları dışında bu çalışmada örneklem büyüklüğünün 3000 olarak sabit tutulması başka bir önemli sınırlıktır. Araştırma kapsamında örneklem büyüklüğü belirlenirken, modellerin doğru parametre kestirimleri sağlamasına yetecek bir örneklem büyüklüğü seçimine dikkat edilmiştir. Ancak literatürde 3000'den daha küçük örneklem büyüklüğü sahip çalışmalara rastlamak oldukça mümkündür (örn., Conway & Huffcutt, 2003; Henson & Roberts, 2006; Jackson, Gillaspy, & Purc-Stephenson, 2009). Bunun yanında 3PL MTK modelini veya DINA modelini kullanmak için gerekli minimum örneklem büyüklüğünün 1000 olması tavsiye edilirken 4PL MTK ile madde parametrelerinin doğru kestiriminde gerekli minimum örneklem büyüklüğüne ilişkin çalışmalara ihtiyaç vardır (de la Torre ve diğerleri, 2010; Hulin ve diğerleri, 1982). Bu doğrultuda gelecek çalışmalarda farklı örneklem büyüklüklerini dikkate alarak 4PL MTK modeli için gerekli minimum örneklem büyüklüğü araştırmanın ve örneklem büyüklüğünün diğer çalışma koşullarıyla etkileşimini incelemenin 4PL MTK ile ilgili literatüre önemli katkılar sağlayacağı düşünülmektedir.