COMPARISON OF IRT MODELS WITH DIFFERENT GUESSING PARAMETERS

FARKLI TAHMİN PARAMETRELERİ İLE MTK MODELLERİNİN KARŞILAŞTIRILMASI

Dr. Öğr. Üyesi Fatih ORÇAN
Trabzon University
Fatih Collage of Education
fatihorcan@trabzon.edu.tr
ORCID: 0000-0003-1727-0456

Received 5 May 2021 - Accepted 4 June 2021
Gönderim 5 Mayıs 2021 - Kabul 4 Haziran 2021

Abstract: ***The Three-Parameter Logistic (3PL) model have some advantages over the other Item Response Theory (IRT) models for multiple-choice testing. Under the 3PL model, an examinee with no knowledge can correctly answer an item at the probability of the value of the c-parameter. The propensity for the guessing effect is the same for all ability levels under 3PL models. However, the idea of ability-based guessing has been asserted. In this study, different IRT models for which the guessing parameters are considered in different ways were elaborated. Also, the IRT models were compared with each other via a simulation study and an empirical data set. The results were compared based on item parameter estimation bias and RMSE. Based on the results, the FG3PL model gave the worst results (i.e., larger bias and RMSE) compared to other models. C3PL model was fine when the simulated data were generated by the 2PL model, but not by 3PL data.***

Keywords: ***Item response theory (IRT), three-parameter logistic (3PL), guessing parameter, ability-based guessing***

Öz: ***Üç Parametreli Lojistik (3PL) modeli, çoktan seçmeli testler için diğer Madde Tepki Teorisi (MTK) modellerine göre bazı avantajlara sahiptir. 3PL modelinde, soru hakkında bilgisi olmayan bir kişi dahi, bir maddeye tahmin parametresi olasılığında doğru cevap verebilir. Bununla birlikte, 3PL modelinde tahmin parametresinin etkisi tüm yetenek seviyeleri için aynıdır. Fakat buna karşın, yetenek temelli tahmin fikri ileri sürülmüştür. Bu çalışmada, tahmin parametrelerinin farklı şekillerde ele alındığı farklı MTK modelleri incelenmiştir. Bu MTK modelleri bir simülasyon çalışması ve gerçek bir veri seti aracılığıyla birbirleriyle karşılaştırılmıştır. Sonuçlar madde parametresi tahmin yanlılığı ve KOKH'ye dayalı olarak karşılaştırıldı. Sonuçlara dayanarak, FG3PL modeli diğer modellere kıyasla en kötü sonuçları (yani daha büyük yanlılık ve KOKH) vermiştir. Fakat, yapay veriler 2PL modeli tarafından oluşturulduğunda C3PL modeli iyiyken 3PL verileri tarafından oluşturulan verilerde iyi değildi.***

Anahtar Kelimeler: ***Madde Tepki Kuramı (MTK), üç-parametre lojisitk (3PL), tahmin parametresi, yetenek temelli tahmin.***

INTRODUCTION

Multiple-choice (MC) is one of the item formats which was preferred by educators for numerous reasons. For instance, MC items are preferred because it can be used in large groups and scored automatically (Roberts, 2006). In addition, the content validity of a test may be greater for a test consisting of MC items (McCoubrie, 2004), because it is possible to include a larger number of test items in a limited testing period. In MC testing, the Three-Parameter Logistic (3PL) Item Response Theory (IRT) models are used instead of One-Parameter Logistic (1PL) and Two-Parameter Logistic (2PL) models, as it is the only IRT model adjusts for the probability of obtaining the correct answer for an item by including the guessing parameter (g), as with the following formula

$$P_j(\theta) = g_j + (1 - g_j)\frac{\exp\left[a_j(\theta - b_j)\right]}{1 + \exp\left[a_j(\theta - b_j)\right]}$$

where θ represents the trait level of an examinee, and $a_j, b_j,$ and g_j represent the item discrimination, difficulty, and guessing (a.k.a., pseudo-guessing) parameters of item *j* respectively. It is believed that guessing behavior happens when the examinee does not know the correct answer for an item but tries to find it randomly (Han, 2012). The guessing parameters are usually smaller than 1/*L*, where *L* is the number of response options (Chalmers, 2012; Embretson & Reise, 2000).

From the perspective of the 3PL model, the probability of success is greater than the *g* value for all ability levels. Also, the propensity to guess is the same for all ability levels. From a different standpoint, the "presence of guessing in the model assumes that, regardless of a person's location, his or her propensity to 'guess' is constant across the continuum" (de Ayala, 2009). From this perspective, the 3PL model does not seem to take ability levels into account when assessing the probability of successful guessing. However, as Embretson and Reise (2000: 71) pointed out, "if examinees can systematically eliminate [an] implausible distractor, selecting the correct answer from the remaining alternatives will have a higher probability than random guessing". Furthermore, "the addition of the pseudo-guessing parameter further increases the potential disordering of the probabilities from ability to ability" (Pelton, 2002: 11). Additionally, the *b* parameter in 3PL is not interpreted as it is according to the 2PL model. Namely, according to 2PL, the item difficulty parameter mostly indicates the location where the probability of success is .5, while this is not the case for the item difficulty parameter according to 3PL. "Item difficulty has a different meaning in the 3PL model" (Embretson & Reise, 2000: 72). Therefore, it is not wrong to state that the inclusion of the guessing parameter changes not only the scoring procedure but also the meaning of the *b* parameters.

The guessing parameter can be conceptualized in two ways (Embretson & Reise, 2000; Gao, 2011; San Martin, del Pino & de Boeck, 2006). In the first case, guessing may be completely random, where responders randomly guess at an answer from the available choices. For instance, respondents may guess at the answer to an item that includes a long reading passage due to time limitations; or an examinee may have no

knowledge about the concept of an item and randomly guess at the answer. In these cases, from a mathematical perspective, the probability of choosing the correct answer is 1/*L*, where *L* is the number of choices. In the second case, guessing may not proceed fully at random; rather, the examinee may apply some information (i.e., partial knowledge) to eliminate some of the answer choices, and then guess from remaining choices. For example, another examinee may eliminate some of the choices for the same item above, leaving only *L** choices (*L**< *L*). Then, out of the *L** choices, the answer may be correctly guessed (assuming one of *L** choices is the correct answer) with a probability of 1/*L**. On the other hand, if an examinee thinks that he or she knows the answer at first hand (whether correct or incorrect), they may not try to guess the answer, but answer the item directly. However, the 3PL model does not distinguish these guessing concepts; and this model also assumes that guessing is a property of the item only (San Martin *et al*., 2006). Considering the cases above, the 3PL model fails to distinguish random guessers and examinees with partial knowledge. Therefore, using the same guessing parameters for all examinees can be considered unfair since "a person guesses on any particular item does not affect estimation of that person's θ" (Waller, 1989: 234).

Burton (2002) discussed the issues of partial knowledge and random guessing in multiple-choice tests (focusing on true-false items) with respect to medical data. Referring to the theoretical distinction between "full knowledge" and "partial knowledge," he emphasized that examinees try to guess an answer when they have no knowledge about the item. Furthermore, Chiu and Camilli (2013) reviewed the literature on the 3PL guessing parameter and discussed some potential issues, asserting that the impact of the guessing parameter on ability estimation is not clear. As they explain, even though the 3PL model produces different probabilities of correctly responding to an item for different ability levels, "students with very low proficiencies have probabilities greater than zero of answering even the most difficult items" (p. 82). On the other hand, as they point out, under the 3PL model, more proficient students receive greater credit in comparison to lower-proficiency students; as such they assert that "3PL scoring raises an equity issue because lower proficiency examinees are not afforded full opportunity to benefit from correctly answering more difficult items" (p. 83).

In an effort to address the guessing issue Han (2012) suggested a redefinition of the problem solving and guessing process. The author offered a logical explanation for the "probability of successful guessing," claiming that successful guessing depends neither on item content nor the ability of the examinees. Rather, it was recommended that the guessing parameter be fixed to 1/*L*, where *L* is the number of choices. Thus, if there were a total of five choices for answering an item, then the guessing parameter would be fixed at .2. This concept is termed the Fixed Guessing Three-Parameter Logistic Model (FG3PL); it is easy to apply in empirical studies since it does not involve additional calculations/parameters to be estimated. The author concluded that FG3PL was superior to 3PL in two ways. First, it had higher parameter estimation convergence. Specifically, "FG3PL turned out to be a much more feasible application even with small sample size" and FG3PL models were more appropriate when there were fewer choices for the MC (p., 14). The second, the FG3PL model had stable and accurate parameter estimation. That is, "FG3PL offered very stable parameter estimation even with smaller sample with moderate

sparseness" (p., 14). However, this procedure starts from the guessing level in calculating the probability of success for all examinees, regardless of their ability level or the item difficulty parameter, as with 3PL. Also, fixing the guessing parameter to a constant might cause estimation problems (Embretson & Reise, 2000).

In search of an ability-based guessing parameter, Gao (2011) introduced a new IRT model (e.g., 2PL-Guessing), which incorporated both item properties and examinees' abilities to estimate the guessing parameter. Along with the general formulation, a few assumptions about the new IRT model were introduced; the proposed guessing parameter will always fall between 1/*L* and .5. The author concluded that the 2PL-Guessing model produced ability estimations that correlated more closely to the true ability of examinees than the 3PL model. Furthermore, the 2PL-Guessing model demonstrated a smaller parameter estimation bias with a higher model-data fit. Thus, from one point of view, 2PL-Guessing serves as an updated version of FG3PL, as in both the 2PL-Guessing and FG3PL models, guessing was related to the number of choices, and the probability of success is higher, as with 3PL. On the other hand, 3PL has been criticized for simply increasing the probability of success. However, the application of 2PL-Guessing models is highly complex and, to the researcher's knowledge, it is not possible to use it in an empirical study with current IRT software.

Under the 3PL model, the estimation of the guessing parameter is unstable (San Martin, del Pino & de Boeck, 2006). As another means to avoid issues with estimation problems in 3PL, constraining the guessing parameter to an estimated common value for all the items in the test (we will call this C3PL for convenience) has also been suggested, thereby reducing the number of estimated parameters (Embretson & Reise, 2000). That is, the C3PL models fixes all the guessing parameters equal to each other. In this sense, the C3PL model reduces complexity. However, it still does not answer the question of whether guessing is ability-related or not.

Even though the 2PL, 3PL, C3PL, and FG3PL models were described in detail they were not compared with each other. Therefore, the comparison of the models is still missing in the literature. So, it is important compare these models to see how they behave under different conditions and which model is superior to the others, if any. Yet, the purpose of this study was to compare the 2PL, 3PL, C3PL, and FG3PL models systematically in terms of item parameter recovery by item estimation bias, Root Mean Squared Errors (RMSE), and the correlation between estimated item parameters and their true values via a simulation study (study 1). Likewise, the models were also compared in terms of item and test information functions by using a real data set (study 2).

1. STUDY 1

1.1.Study 1 Method

In order to compare the models, 500 data sets were generated by WinGen 3 (Han, 2007) for each conditions. For that, the ability parameters of the examinees were first generated from a normal distribution with a mean of 0 and a standard deviation of 1

($\theta \sim N(0,1)$). According to Gao (2011), when the focus of a study is on parameter estimation values, the simulation should be repeated 500 times, therefore; in this study, 500 repetitions were performed for each simulation condition, as specified below. By using the θ values, the examinees' responses were generated for pre-specified item parameters (see Table 1) with respect to the 2PL and 3PL models. Then, the responses were analyzed for all 2PL, 3PL, C3PL and FG3PL models by using IRTPro 2.1 (Cai, L., Thissen, D., & du Toit, S. H. C., 2011) software. Based on the results, the correlation between the estimated item parameters and their true values were first calculated. Then, the absolute relative bias and the Root Mean Squared Errors (RMSE) for item parameter estimation were calculated for all four models. It was expected that values of the parameter estimation bias would be smaller than .05 for a good result (Hoogland & Boomsma, 1998). Similarly, the RMSE shows the variation among the estimated values of a parameter. Therefore, the smaller the RMSE, the more accurate the estimation results (Gao, 2011). Afterward, the ability parameter estimations for all four models were compared according to the item characteristics curves (ICC), the item information function (IIF) and test information function (TIF).

The bias was calculated using:

$$Bias = \frac{|\hat{k} - k|}{k}$$

where $\hat{k}$ and k represent the mean of the parameter estimate and the true value of the parameters, respectively. The RMSE was calculated by:

$$RMSE = \sqrt{\frac{\sum(\hat{k_i} - k)^2}{n}}$$

where n shows the number of replication and $\hat{k_i}$ shows the estimated value of replication i.

1.2.Simulation Design Factors

The design factors which were commonly used in IRT studies were considered for the current study.

- Data generation model: Two different models were considered for data generation; 2PL and 3PL models.
- Analyzing model: The simulated data were analyzed with four different models to estimated item parameters; 2PL, 3PL, C3PL and FG3PL.
- Sample size: Three different sample sizes were considered with this study; 1200, 2500, and 4000. In order for an accurate parameter estimated under 3PL model the sample size was suggested to be larger than 1000 (Finch & French, 2019; Lord, 1968). To be more conservative, the sample sizes were considered above the suggested value.
- Number of item: For this study two different number of items were considered; 20 and 30 items. Swaminathan and Gifford (1983) suggested at least 20 items for a problem free estimation with sample size of 1000 under

3PL model (as cited in Akour & Al-Omari, 2013). Therefore, item sizes were chosen above suggested values in order to not to have any problems related to it.

- Item parameter values: True item parameters were reported at Table 1. There were only 20 items shown in the table. If the number of items was 30 the same item parameters for first 10 items were set for items 21 through 30.

Table 1: True Item Parameters

	3PL		
	2PL		
Items	A	b	g
1	1.37	-.17	.1
2	.90	.55	.3
3	.43	2.31	.1
4	1.00	-.53	.3
5	.65	.71	.1
6	.31	-.05	.3
7	1.16	-.15	.1
8	.40	-1.09	.3
9	1.08	1.62	.1
10	1.01	-0.04	.3
11	1.23	.44	.1
12	.92	.49	.3
13	.62	-.12	.1
14	1.43	-.56	.3
15	1.31	1.04	.1
16	1.16	.71	.3
17	.58	.38	.1
18	.97	-2.03	.3
19	.83	-.65	.1
20	1.25	-.65	.3

The item parameters were generated with truncated normal distribution as in Paek (2014). For item discrimination parameter the mean was set to .9, standard deviation was .2, the minimum was .3 and the maximum was 1.5. For the difficulty parameters the values were set to 0, 1, -3, and 3, respectively. Thus, some of the items has low discrimination (i.e., a <.6) while others have medium and large discrimination (i.e., a >1.2) values. Similarly, the difficulty parameters ranged from small to large values to represent easy and difficulty items. For the guessing parameters values were set to .1 or .3 to represent low and high guessing values.

1.3. Study 2 Results

The purpose of the first study was to compare the item parameter recovery between 2PL, 3PL, C3PL and FG3PL models via a simulation study. For that, item parameter bias and RMSE values were calculated via the formulas given above. Moreover, the average item parameter estimation values were correlated with their true values. The correlation between the parameter estimates were reported at Table 2. The values were ranging between .574 and 1.000. When the data were generated by 2PL and analyzed with the same model the correlation values were all 1.000 for both *a* and *b* parameters, regardless of the number of items and the sample size. However, the correlations get smaller when the data were generated with 3PL and analyzed with 3PL model. On the other hand, when the sample size was increased

the correlation values get larger. For example, if the sample size (SS) was 1200 and the number of item (NI) was 20 the correlation coefficient was .574 for the *a* parameter under 3PL. When the SS was increased to 2500 the correlation became .979. Changing the number of items was effective only on the *a* parameter under the small sample size and 3PL. However, the change was small for the *b* parameters, which was only at third decimal.

Table 2: Correlation Between True and Estimated Item Parameters

		20 items				30 items			
		2PL		3PL		2PL		3PL	
		a	b	a	b	a	b	a	b
1200	2PL	1.000	1.000	.931	.870	1.000	1.000	.941	.860
	3PL	.990	.977	.574	.991	.997	.977	.719	.992
	C3PL	.996	1.000	.890	.896	.998	1.000	.917	.918
	FG3PL	.743	.956	.782	.941	.746	.951	.810	.943
2500	2PL	1.000	1.000	.922	.871	1.000	1.000	.934	.870
	3PL	.997	.990	.979	.992	.999	.990	.950	.994
	C3PL	1.000	1.000	.924	.919	1.000	1.000	.940	.918
	FG3PL	.783	.960	.818	.948	.775	.952	.837	.950
4000	2PL	1.000	1.000	.925	.876	1.000	1.000	.938	.873
	3PL	.999	.987	.996	.998	.999	.988	.997	.998
	C3PL	.999	1.000	.894	.928	.999	1.000	.920	.925
	FG3PL	.761	.960	.785	.948	.760	.951	.813	.949

For C3PL model in which the guessing parameters of the items were fixed to each other, the correlation for *a* and *b* parameter ranged between .890 and 1.000. The correlations under C3PL model were higher for the data generated from 2PL compared to 3PL model. For example, for SS = 1200 and NI = 20, the correlation of C3PL model under 2PL data was .996 while it was .890 under 3PL model for the *a* parameter. For the data generated with 3PL model and SS = 1200 the correlations for C3PL models (r=.890) bigger than the corresponding value under 3PL model (r=.574). However, as the sample size was increased the gap between them had decreased. On the other hand, for the data generated under 2PL model, there were not much difference among the models by the SS.

Under FG3PL model where the guessing parameters were fixed at .20, the correlations were between .743 and .960. Increasing the sample size or number of items did not affect the correlation coefficients. However, the correlations were generally smaller compared to the other model. The differences were more obvious for the *a* parameter, while it was around %5 for the *b* parameter. Besides, the correlations from 3PL data were bigger than 2PL data.

Table 3 and 4 show maximum, minimum and mean values of the item parameter estimation bias and RMSE for the 2PL and 3PL data sets. Based on the results, the number of items did not have much effect on the average bias of the items. Similar conclusions can also be made for the maximum and minimum values. When the sample size was small and the data were generated with 2PL model, the biases of the *a* parameters were smaller for 2PL and C3PL models compared to the other model

values and increasing the SS only affected the 3PL and FG3PL model parameters. However, for the *b* parameter, increasing the sample size did not alter average bias much, neither for 3PL nor for FG3PL. In general, for the *a* and *b* parameters average biases were around 8% for 3PL model and around 50% for FG3PL model. Consequently, the 3PL and FC3PL models produced much higher bias values than the suggested level for the 2PL data.

Table 3: Parameter Estimation Under 2PL Data

			20 items						30 items					
			Bias			RMSE			Bias			RMSE		
		Model	Max	Min	Mean	Max	Min	Mean	Max	Min	Mean	Max	Min	Mean
a	1200	*2PL*	.02	-.02	.01	.12	.07	.10	.02	-.01	.01	.11	.06	.09
		3PL	.33	.08	.14	3.12	.19	.39	.25	.08	.13	.90	.19	.26
		C3PL	.11	-.02	.02	.18	.07	.11	.09	-.01	.02	.17	.06	.10
		FG3PL	2.17	.07	.56	2.24	.12	.60	2.06	.06	.55	2.15	.12	.58
	2500	*2PL*	-.01	-.05	.02	.10	.05	.07	-.01	-.04	.02	.09	.05	.07
		3PL	.14	.00	.04	.25	.10	.13	.13	.01	.04	.23	.10	.12
		C3PL	-.01	-.03	.02	.10	.05	.07	-.01	-.03	.02	.09	.05	.07
		FG3PL	1.84	.08	.54	1.88	.10	.51	1.77	.06	.49	1.80	.09	.50
	4000	*2PL*	.03	.00	.01	.07	.04	.05	.03	.01	.01	.07	.04	.05
		3PL	.13	.06	.08	.19	.11	.13	.12	.05	.08	.18	.10	.12
		C3PL	.08	.01	.03	.11	.04	.06	.07	.01	.02	.10	.04	.06
		FG3PL	2.13	.10	.59	2.15	.11	.60	2.00	.09	.58	2.02	.11	.59
b	1200	*2PL*	.06	-.10	.03	.44	.06	.13	.06	-.08	.03	.54	.06	.14
		3PL	.76	-.13	.14	1.40	.10	.38	.66	-.21	.15	1.34	.10	.40
		C3PL	.09	-.07	.02	.45	.06	.13	.08	-.07	.02	.54	.05	.14
		FG3PL	1.34	.12	.49	1.40	.17	.52	1.29	.08	.49	1.33	.15	.53
	2500	*2PL*	.07	-.08	.03	.31	.05	.09	.10	-.08	.04	.35	.04	.10
		3PL	.46	-.10	.08	1.16	.07	.29	.43	-.13	.09	1.13	.09	.32
		C3PL	.08	-.07	.03	.31	.04	.09	.11	-.07	.03	.35	.04	.10
		FG3PL	1.33	.16	.50	1.35	.23	.52	1.31	.12	.51	1.33	.18	.52
	4000	*2PL*	.04	-.01	.02	.22	.04	.07	.04	-.03	.02	.23	.03	.07
		3PL	.57	-.10	.13	1.15	.05	.26	.49	-.13	.12	1.11	.05	.27
		C3PL	.07	-.02	.04	.22	.04	.08	.06	-.02	.03	.23	.04	.08
		FG3PL	1.29	.07	.52	1.31	.15	.53	1.28	.01	.52	1.29	.12	.54
g	1200	*3PL*	.18	.01	.07	.29	.02	.11	.21	.01	.07	.31	.02	.11
		C3PL	.01	.01	.01	.02	.02	.02	.01	.01	.01	.01	.01	.01
	2500	*3PL*	.14	.01	.04	.23	.01	.08	.13	.01	.04	.21	.01	.07
		C3PL	.00	.00	.00	.01	.01	.01	.00	.00	.00	.00	.00	.00
	4000	*3PL*	.13	.01	.04	.20	.01	.07	.12	.01	.04	.19	.01	.07
		C3PL	.01	.01	.01	.01	.01	.01	.01	.01	.01	.01	.01	.01

Note: The true *g* parameter for the data generated with 2PL model was set to zero.

The guessing parameters (*g*) for FG3PL model were not reported since the values were all fixed. For the data generated under 2PL the g parameters of the models were not dependent number of items. The average biases were all smaller than %10. Increasing the sample size also increased the *g* parameter for 3PL but did not change it for the C3PL model. When the data generated with 3PL model, as the sample size increased the average bias of the *a* parameter decreased only for the 3PL model. That is, the biases were not changed for the 2PL, C3PL and FG3PL. Besides, increasing

the number of item did not alter the average bias. Similar interpretations can also be made for the RMSE values. For example, as the sample size increased average RMSE for the *a* parameter remarkably decreased for the 3PL model.

Table 4: Parameter Estimation Under 3PL Data

			20 items						30 items					
			Bias			RMSE			Bias			RMSE		
		Model	Max	Min	Mean	Max	Min	Mean	Max	Min	Mean	Max	Min	Mean
a	1200	*2PL*	-.06	-.50	.20	.50	.10	.23	-.06	-.49	.19	.50	.10	.22
		3PL	1.36	.09	.23	1.09	.33	1.09	1.19	.08	.22	8.49	.27	1.06
		C3PL	.28	-.36	.14	.41	.10	.20	.21	-.37	.12	.38	.09	.18
		FG3PL	.97	-.25	.22	1.06	.09	.28	.89	-.25	.21	.96	.09	.26
	2500	*2PL*	-.07	-.52	.22	.52	.09	.24	-.07	-.52	.21	.52	.09	.22
		3PL	.22	-.09	.06	.52	.17	.26	.52	-.08	.07	3.73	.14	.37
		C3PL	.01	-.42	.11	.43	.06	.16	-.01	-.42	.10	.43	.06	.15
		FG3PL	.75	-.29	.20	.79	.08	.23	.72	-.29	.19	.75	.07	.22
	4000	*2PL*	-.05	-.50	.20	.50	.06	.21	-.05	-.49	.19	.50	.07	.20
		3PL	.18	.04	.08	.31	.15	.22	.19	.05	.08	.32	.14	.20
		C3PL	.24	-.37	.13	.37	.06	.16	.18	-.37	.12	.38	.06	.14
		FG3PL	.96	-.25	.23	.98	.05	.25	.88	-.24	.22	.89	.05	.23
b	1200	*2PL*	-.12	-2.93	.78	3.92	.22	.88	-.12	-3.05	.84	3.81	.21	.93
		3PL	.27	-.31	.12	1.95	.15	.64	.37	-.35	.13	2.05	.15	.68
		C3PL	.20	-2.34	.44	12.81	.09	1.06	.18	-1.88	.46	3.52	.08	.61
		FG3PL	.39	-1.22	.37	3.23	.14	.53	.40	-1.21	.38	3.66	.14	.53
	2500	*2PL*	-.03	-2.93	.77	3.04	.15	.80	-.03	-2.88	.82	2.97	.15	.86
		3PL	.17	-.30	.13	2.02	.11	.57	.14	-.35	.15	1.99	.11	.59
		C3PL	.16	-1.93	.43	2.03	.06	.50	.15	-1.95	.49	2.02	.06	.55
		FG3PL	.41	-1.05	.36	1.11	.17	.39	.41	-1.04	.38	1.08	.16	.41
	4000	*2PL*	-.10	-2.68	.71	2.73	.14	.73	-.10	-2.68	.76	2.72	.14	.78
		3PL	.19	-.21	.07	1.90	.07	.47	.30	-.23	.06	1.85	.07	.49
		C3PL	.14	-1.59	.38	1.62	.07	.41	.10	-1.63	.41	1.67	.06	.45
		FG3PL	.42	-.94	.33	.97	.09	.35	.41	-.93	.34	.95	.09	.36
g	1200	*3PL*	.07	-.04	.03	.32	.05	.16	.07	-.04	.02	.32	.05	.15
		C3PL	.03	-.17	.10	.17	.04	.11	.03	-.17	.10	.17	.03	.10
	2500	*3PL*	.02	-.06	.03	.29	.05	.14	.01	-.06	.03	.27	.04	.13
		C3PL	.01	-.19	.10	.19	.02	.11	.00	-.20	.10	.20	.02	.11
	4000	*3PL*	.03	-.03	.01	.26	.03	.12	.03	-.02	.01	.25	.03	.11
		C3PL	.03	-.17	.10	.17	.03	.10	.02	-.18	.10	.18	.02	.10

However, the changes were between .01 and .04 for the other models. Number of item was not a factor for the *b* parameter estimations either; average bias values were comparable with RMSE values. However, increasing sample size decreased average RMSE values for b parameter. Last but not least, increasing sample size also decreased the values of RMSE for the *g* parameter under 3PL but not under C3PL.

2. STUDY 2

2.2. Study 2 Method

In order to compare the 2PL, 3PL, C3PL and FG3PL models in terms of ICCs and TIFs, a data set from an online university history course was analyzed. The data, which comprised 4000 student responses for 24 dichotomously scored items. The average number correct score for the students was 11 questions. 48% of the students scored higher than the average number correct score. The data were analyzed by the IRTPro 2.1 program (Cai, L., Thissen, D., & du Toit, S. H. C., 2011) in order to estimate the parameters under the models.

2.3.Study 2 Results

First, the correlations among the ability estimations were reported in Table 5. The correlation coefficients ranged between .947 and .999, and all were significant at .01 alpha level. The minimum correlation was between the 3PL and C3PL model, while the maximum was between the C3PL and FG3PL models. The values of the skewness and kurtosis were also reported Table 5. Besides, the distributions of the ability parameters were reported in Figure 1. The shapes of the ability distributions were alike. All of the skewness values were positively, while FG3PL had the most skewed distribution.

Table 5: The Correlation among Ability Estimations

	2PL	3PL	C3PL	FG3PL	Skewness	Kurtosis
2PL	1.00				.276	-.313
3PL	.950*	1.00			.306	-.442
C3PL	.991*	.947*	1.00		.324	-.380
FG3PL	.987*	.950*	.999*	1.00	.352	-.407

*: $p < .01$

Also, the TIFs were reported in Figure 2. When the abilities were negative, the 2PL model was more informative than the other models. However, when the abilities were positive, the 3PL model generally gave the most information. Based on the results pictured at Figure 2, C3PL and FG3PL models produced very close test information functions.

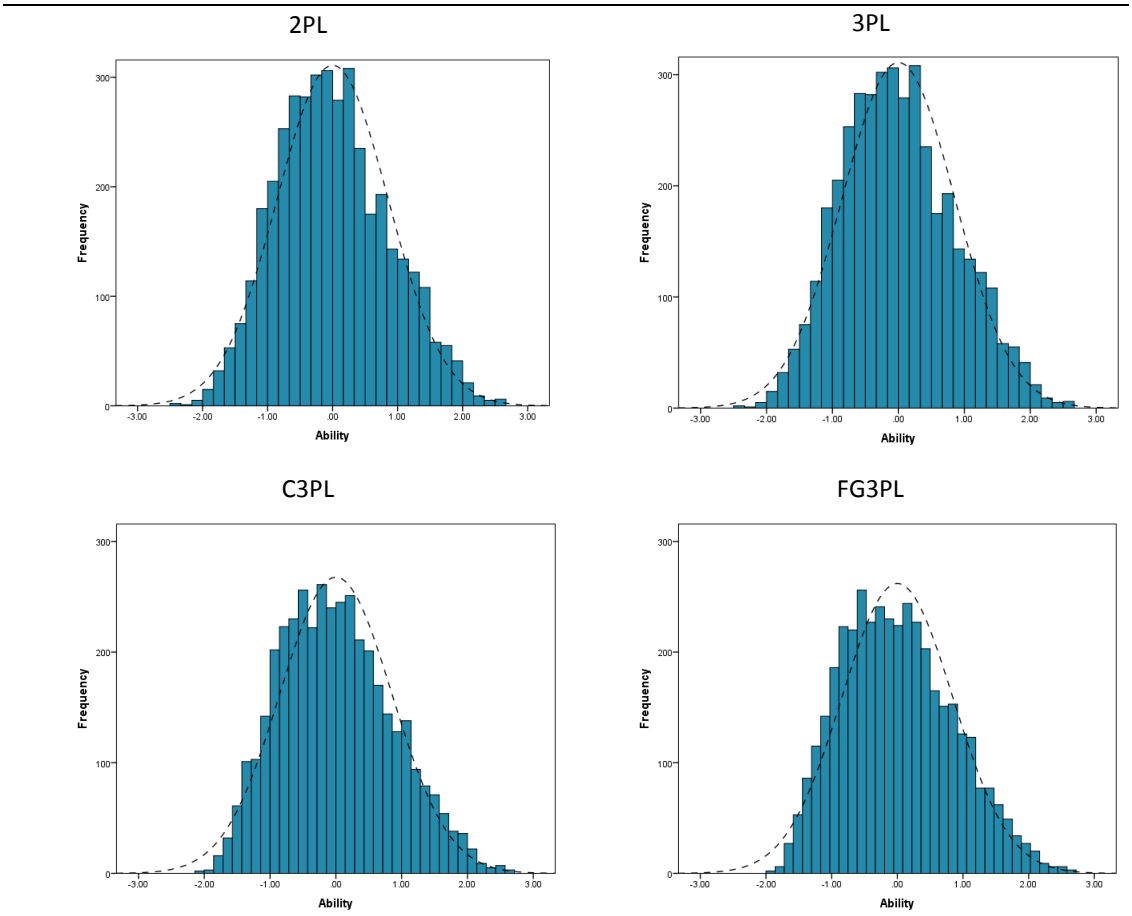

Figure 1: Distributions for the Ability Estimation

The ICC and IIF were provided only for two items in Figure 3 due to space limitations. For both items the 3PL model was the most informative for most of the abilities. Under the second item, FG3PL was the most informative for some of the positive ability levels. As it was the case for TIF, C3PL and FG3PL produced similar information lines under item 1. At the right side of Figure 3, the ICC of the items were shown. For the second item, the three parameter models, showed similar curves. That is, 3PL, C3PL and FG3PL models showed almost the same ICCs. A different pattern can be seen for the first item given at Figure 3. The difference was due to the discrepancies for lower ability levels. Thus, there exist small differences at the lower abilities.

Finally, the models' fit were compared with -2*loglikelihood values (chi-square). Among all four models 3PL model showed larger likelihood value (-2*loglikelihood=123537.7). The chi-square difference tests were used to test the difference between the models. Based on the results, 3PL model showed significantly better fit compared to 2PL, C3PL and FG3PL models $\Delta\chi^2{}_{3PL-2PL} = 55.90,\ \Delta df = 24;\ \Delta\chi^2{}_{3PL-C3PL} = 209.15,\ \Delta df = 23;\ \Delta\chi^2{}_{3PL-FG3PL} = 177.34,\ \Delta df = 24)$

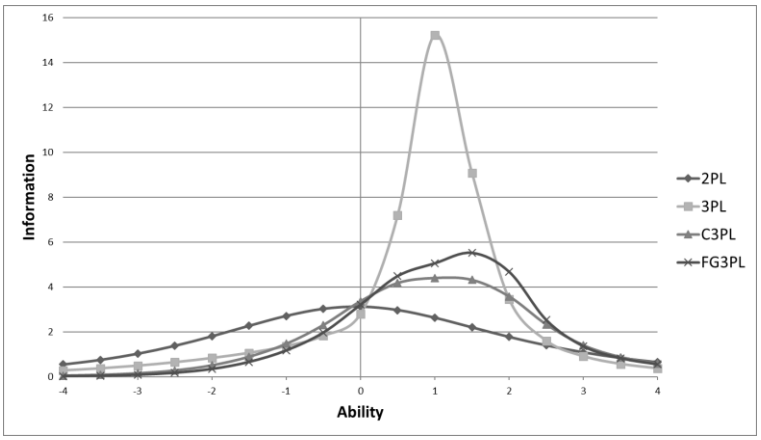

Figure 2.Test Information Functions

DISCUSSION AND CONCLUSION

In this study, the literature on ability-based guessing was reviewed and simulation analysis were conducted to compare different three parameter models (3PL, C3PL and FG3PL). Based on the result of the simulation study, the data generated and analyzed under 3PL model, samples size had an effect on the parameter estimates while there was not such an effect under the data set generated and analyzed with 2PL model. As it was pointed in the literature, this results showed that 3PL model requires lager sample sizes for consistent parameter estimation compared to 2PL models (de la Torre & Hong, 2010; de Ayala, 2009). On the other hand, increasing the SS made small changes on the correlations between the item parameters under C3PL and FG3PL models. As reported at Table 2, the lowest correlations between the true and estimated parameter values mostly belong to the FG3PL model, especially for the *a* parameter. Therefore, fixing the guessing parameter to 1/L was not rational, especially in terms of the *a* parameter. Similar comments can also be made for C3PL model. However, the correlations values were higher for the C3PL model compared to the FG3PL model. Also, when the data were generated with no guessing values (a.k.a., 2PL model) 2PL, 3PL and C3PL model gave almost perfect correlation for both parameters. However, since FG3PL model fixed the guessing parameter as 1/*L*, the correlation values were small for both parameters. That is, with or without guessing values FG3PL model gave the worst result compared to the other models.

Even though Han (2012) suggested that FG3PL model was stable with small SS, the current study showed not much difference between the large and small SS. This might be due to the SS used in this study which were larger than the SS used in Han's (2012) study. Specifically, when the data were generated with 2PL model and sample size was small the average bias and RMSE values were smaller for 2PL and C3PL model while they were higher for 3PL and FG3PL models. The biases were smaller than 5% for 2PL and C3PL. However, the results were different for the data sets generated by 3PL model. Although, under small sample size the models were not different much, when sample size was increased, average bias and RMSE values were smaller for 3PL model compared to the other models with 3PL generated data sets. Under this circumstance, neither C3PL nor FG3PL gave good results. For these

model the biases of the *b* parameter were around 35% even when the sample size was 4000. At same condition the bias was around 70% for 2PL model. On the other hand, the bias values for the *a* parameter were smaller; however, they were still larger than %10 critical level. That is to say, the FG3PL model did not give a good result for none of 2PL and 3PL data. Namely, fixing the guessing parameter to inverse of number of choices distorted the parameter estimations. Conversely, the C3PL models resulted well in terms of item parameter estimations under 2PL data. However, under 3PL model the results were far from the true values. This must be due to the way that C3PL was defined. That is, data generated under 2PL model had guessing values set to zero. Therefore, with C3PL model estimated guessing values were all equal and they were close to zero which were the true values. Thus, the bias values were small. When it comes to 3PLgenerated data fixing the guessing values equal mislead the overall parameter estimation. To conclude, based on these results, it is not recommended to use C3PL or FG3PL models.

Under the real data analysis, the models were compared using a real data set consisting of 24 items. The ability parameters were estimated with these models, and the correlations between the estimates were calculated. The correlations for abilities were all above .947 and significant. Besides, the shapes of the estimated ability parameters were not much different; while the 2PL-estimated thetas were more normally distributed based on the skewness and kurtosis values. Thus, estimated theta values were alike each other. From this point of view, the model used for theta estimation does not seem to affect the results.

Based on the ICC for given two items, 2PL models differs from other models. In other words, ICCs were not different for 3PL, C3PL and FG3PL models. There seem to be small differences on the lower ability estimates. This must be due to use of different guessing parameters for the 3PL, C3PL and FG3PL. When it comes to the IIF and TIF, C3PL and FG3PL models were not distinguished from each other and do not show better results compared to 3PL model. In summary, 3PL model did not produced worse results than other models which include the guessing effects. In fact, when the guessing effect was present in the data set (i.e., 3PL data), 3PL model produced smaller biases. Therefore, as a conclusion it can be said that C3PL and FG3PL models produced similar results and 3PL model can be chosen over C3PL and FG3PL models. This finding was also supported by the chi-square difference test. Based on the test results, C3PL and FG3PL did not showed better fit than 3PL model. Therefore, it can be concluded that there is no need to use C3PL and FG3PL models instead of a 3PL model.

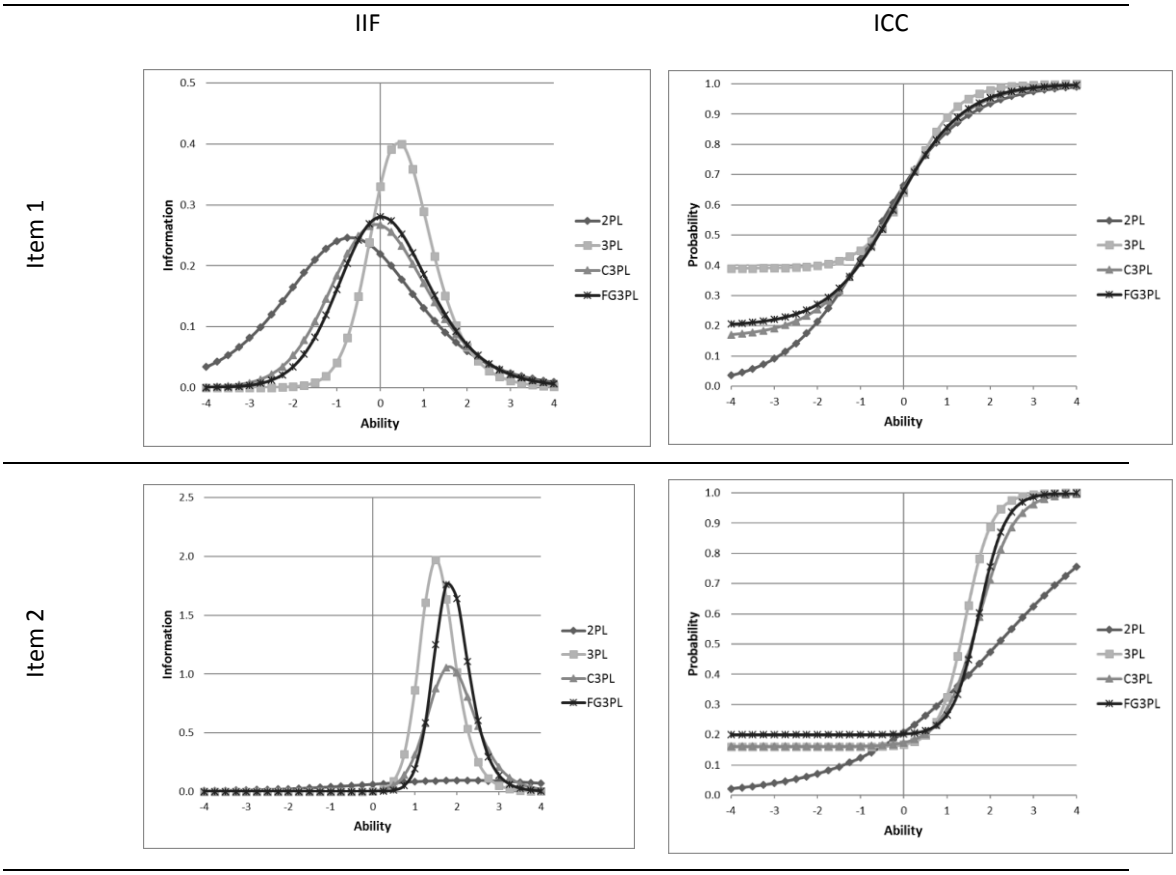

Figure 3: IIFs and ICCs for Two Items

Note: All the item parameters and more figures can be provided by an email if requested from the corresponding author.

The results indicated that alternative 3PL models (i.e., C3PL and FG3PL) does not improve the results compared to actual 3PL model. They generally produced larger bias and RMSE. Also, the alternative models still do not answer the problem of ability based guessing. Therefore, there is still need for a model which considers guessing parameter based on the abilities of the examinees. Only when the consistent guessing effect for high and low ability levels can be controlled. Besides, the results were based on relatively large sample sizes (1200 and higher). C3PL and FG3PL models had less parameter to estimate compared to a 3PL model. Therefore, the results might be different small sample sizes. To explore it, another simulation study with smaller sample size would be useful. Similarly, number of items did have effect on the estimation for none of the models. However, number of items lower than the suggested values may show different results. So, another simulation study with smaller number of items would be useful, too.

REFERENCES

Akour, M. & Al-Omari, H. (2013), "Empirical investigation of the stability of IRT item-parameters estimation", *International Online Journal of Educational Sciences,* 5(2): 291-301.

Burton, R. F. (2002), "Misinformation, partial knowledge and guessing in true ⁄ false tests", *Medical Education,* 36: 805-811.

Cai, L., Thissen, D., & du Toit, S. H. C. (2011), IRTPRO for windows [Computer software]. Lincolnwood, IL: Scientific Software International.

Chalmers, R., P., (2012), "mirt: A Multidimensional item response theory package for the R environment", *Journal of Statistical Software,* 48(6): 1-29.

Chiu, T. & Camilli, G. (2013), "Comment on 3PL adjustment for guessing", *Applied Psychological Measurement,* 37(1): 76-86.

de la Torre, J. & Hong, Y., (2010), "Parameter estimation with small sample size a higher-order IRT model approach", *Applied Psychological Measurement, 34*(4): 267 - 285. DOI: 10.1177/0146621608329501.

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

Embretson, S. E., & Reise, S. P. (2000), *Item response theory for psychologists.* Mahwah, NJ: Lawrence Erlbaum.

Finch, H., & French, B. F. (2019), A comparison of estimation techniques for IRT models with small samples. *Applied Measurement in Education, 32*(2): 77-96. DOI: 10.1080/08957347.2019.1577243

Gao S. (2011), "The exploration of the relationship between guessing and latent ability in IRT models", PhD thesis, Southern Illinois University at Carbondale, Department of Educational Psychology and Special Education.

Han, K. T. (2007), "WinGen: Windows software that generates IRT parameters and item responses", *Applied Psychological Measurement*, 31(5): 457-459.

Han, K. T. (2012), Fixing the c parameter in the three-parameter logistic model. *Practical Assessment Research & Evaluation*, 17(1): 1-23.

Hoogland, J. J. & Boomsma, A. (1998), Robustness studies in covariance structure modeling: An overview and a meta-analysis, *Sociological Methods & Research, 26:* 329-367.

Lord, F. M. (1968), An analysis of the verbal scholastic aptitude test using Birnbaum's three- parameter logistic model. *Educational and Psychological Measurement, 28,* 989 – 1020. doi:10.1177/001316446802800401.

McCoubrie, P., (2004), "Improving the fairness of multiple-choice questions: a literature review", *Medical Teacher, 26*:8, 709-712, DOI: 10.1080/01421590400013495

Paek, I., (2014), "An investigation of the impact of guessing on coefficient α and reliability", *Applied Psychological Measurement*, 1-14. DOI: 10.1177/0146621614559516

Pelton, T., W., (2002), "The accuracy of unidimensional measurement models in the presence of deviations for the underlying assumptions", PhD Theses, Brigham Young University, Department of Instructional Psychology and Technology.

San Martin, E., del Pino, G., & de Boeck, P. (2006), "IRT models for ability-based guessing", *Applied Psychological Measurement,* 30(3): 183-203.

Swaminathan, H., & Gifford, J. A. (1983), Estimation of parameters in the three-parameter latent trait model. In D. J. Weiss (Ed.), *New horizons in testing*, (pp. 9-30). New York: Academic Press.

Waller, M.I., (1989), "Modeling guessing behavior: A comparison of two IRT models", *Applied Psychological Measurement. 13*(3): 233-243.