JP3I (Jurnal Pengukuran Psikologi dan Pendidikan Indonesia), 12(2), 2023, 177-190
DOI: http://dx.doi.org/10.15408/jp3i.v12i2.34130
http://journal.uinjkt.ac.id/index.php/jp3i

Ensuring Parameter Estimation Accuracy in 3PL IRT Modeling: The Role of Test Length and Sample Size

Hasan Djidu[1,2], Heri Retnawati[3], Haryanto[4]

Educational Research and Evaluation, Graduated School, Universitas Negeri Yogyakarta, Indonesia[1]
Mathematics Education, Faculty of Teacher Training and Education, Universitas Sembilanbelas November Kolaka, Indonesia[2]
Electrical Engineering, Faculty of Engineering, Universitas Negeri Yogyakarta, Indonesia[3]
Mathematics Education, Faculty of Mathematics and Natural Science, Universitas Negeri Yogyakarta, Indonesia[4]

hasandjidu.2021@student.uny.ac.id, hasandjidu@usn.ac.id

Abstract

The objective of this simulation study was to evaluate the accuracy of item parameters estimation when employing the 3PL IRT model, mainly focusing on sample size and the length of the test (number of test items). The investigation used six datasets produced by WinGen, each comprising 5000 responses and varying test lengths within 10 to 40 items. For each dataset, the study conducted simulations and re-analyzed the data 15 times, generating a total of 2025 data subsets and estimating 225 parameters for each item. The results revealed that smaller sample sizes led to more pronounced biases, emphasizing a recommended minimum sample size of 3000 for precise parameter estimation. Additionally, the study found that a limited number of items (short test) yielded biased estimations and proposed a minimum of 25 or 40 test items for accurate estimation using the 3PL IRT model. These findings offer valuable insights for test developers in making informed decisions regarding sample sizes and test length, ultimately ensuring reliable and accurate parameter estimates.

Keywords: parameter estimation, item response theory, 3-parameter logistics, test length, sample size

Abstrak

Tujuan dari studi simulasi ini adalah untuk mengevaluasi akurasi estimasi parameter item ketika menggunakan model IRT 3PL, dengan fokus utama terkait ukuran sampel dan panjang tes (jumlah item tes). Penelitian ini menggunakan enam set data yang dihasilkan dari WinGen, masing-masing terdiri dari 5000 respons dan beragam panjang tes antara 10 hingga 40 item. Simulasi dilakukan dengan menganalisis ulang setiap dataset sebanyak 15 kali, menghasilkan total 2025 subset data dan mengestimasi 225 parameter untuk setiap item. Hasil simulasi menunjukkan bahwa ukuran sampel yang lebih kecil menghasilkan bias, dengan penekanan pada ukuran sampel minimum yang direkomendasikan sebesar 3000 untuk estimasi parameter yang akurat. Selain itu, studi ini menemukan bahwa jumlah item yang terbatas menghasilkan estimasi parameter item yang bias dan mengusulkan minimum 25 atau 40 item tes untuk mendapatkan estimasi yang akurat menggunakan model IRT 3PL. Temuan ini memberikan wawasan berharga bagi pengembang tes dalam membuat keputusan mengenai ukuran sampel dan panjang tes, yang pada akhirnya memastikan estimasi parameter yang dapat diandalkan dan akurat.

Kata kunci: *estimasi parameter, teori respon butir, 3 parameter logistik, panjang tes, ukuran sampel*

Introduction

One of the crucial assumptions in the item response theory (IRT) is the requirement for parameter invariance. In IRT, parameters refer to both those associated with the test item and those related to the test taker (Paek et al., 2021). "Invariant" signifies that the characteristics of the test item parameters remain unaffected by the test taker's ability, and vice versa; the ability of the test taker is independent of the parameters of the test item (Baker, 1985; Retnawati, 2014; Stenbeck et al., 1992). This invariance of parameters is a fundamental aspect that distinguishes IRT from classical test theory (CTT). Ensuring compliance with this assumption is essential before further analyzing the test items.

The fulfilment of assumptions in IRT models heavily depends on the quality of the test items. Evaluating the quality of these items can be approached from two perspectives: content-wise and statistically (Paek et al., 2021). For test developers, a solid understanding of item development is crucial in crafting test items with high content quality. This ensures that the items effectively assess the intended constructs or latent attributes.

On the other hand, the parameters of the test items obtained from the IRT analysis, such as difficulty (b), discrimination (a), and pseudo-guessing (c), can be statistically evaluated for quality if the underlying model has successfully met the assumptions. The IRT model's assumptions, including parameter invariance and other statistical requirements, play a significant role in determining the quality and validity of the parameter estimates. Overall, a comprehensive assessment of test item quality involves content evaluation during item development and rigorous statistical evaluation using appropriate IRT models. By combining these perspectives, test developers can create more reliable and valid measuring instruments, enhancing the accuracy and effectiveness of the assessment process.

Research Results Related to the Accuracy of Item Parameters and Sample Size

Achieving parameter invariance in IRT poses a significant challenge, especially when dealing with limited sample sizes in pilot studies. The issue of constrained sample size is a common hurdle during the calibration of test items using the IRT approach. Barnes and Wise (1991) conducted a study to explore the accuracy of item parameters with smaller sample sizes. Their findings indicated that enhancing the stability and accuracy of item parameter estimation (e.g., difficulty level and ability) could be achieved by modifying the 1-parameter IRT model. This modification involved setting a specific value for the pseudo-guessing (c) parameter. Similarly, other studies, such as those by Wainer & Wright (1980) and Divgi (1984), have investigated and demonstrated the accuracy of parameters by adopting a modified 1-parameter IRT approach, often referred to as the 'Rasch' model.

These studies underscore that the challenge of sample size affecting item parameter accuracy is a long-standing concern. However, due to the absence of a definitive guide on sample size requirements for calibrating item parameters, ongoing research on sample size in IRT remains a critical consideration for instrument developers. By continuously exploring this aspect, researchers can make informed choices to enhance the reliability and validity of their measurement instruments.

Recent studies, such as the one by Paek et al. (2021), suggest a sample size of at least 500 for calibration using the Rasch model. For longer tests (test length = 40), they recommend a sample size of at least 750. In their research, they conducted simulations employing both the Rasch and IRT 2 PL models, varying sample sizes across 13 scenarios, test lengths from 9 to 40, and different estimation methods (joint maximum likelihood (JML), marginal maximum likelihood (MML), and conditional maximum likelihood (CML)). The results demonstrated that the IRT 2PL model was more sensitive to changes in sample size than the Rasch model. However, it's worth noting that the study by Paek et al. (2021) did not include the use of the 3PL model in their simulations.

Another study by Feuerstahler (2022) recommended a sample size of at least 5000 to achieve the best calibration results with the 3PL model. This suggests that meeting the 3PL model's parameter requirements

http://journal.uinjkt.ac.id/index.php/jp3i

necessitates a large sample size, which can be challenging in practical contexts. On another note, Yen (1981) highlighted the importance of using the 3PL IRT model, particularly for estimating ability, especially in the context of multiple-choice tests.

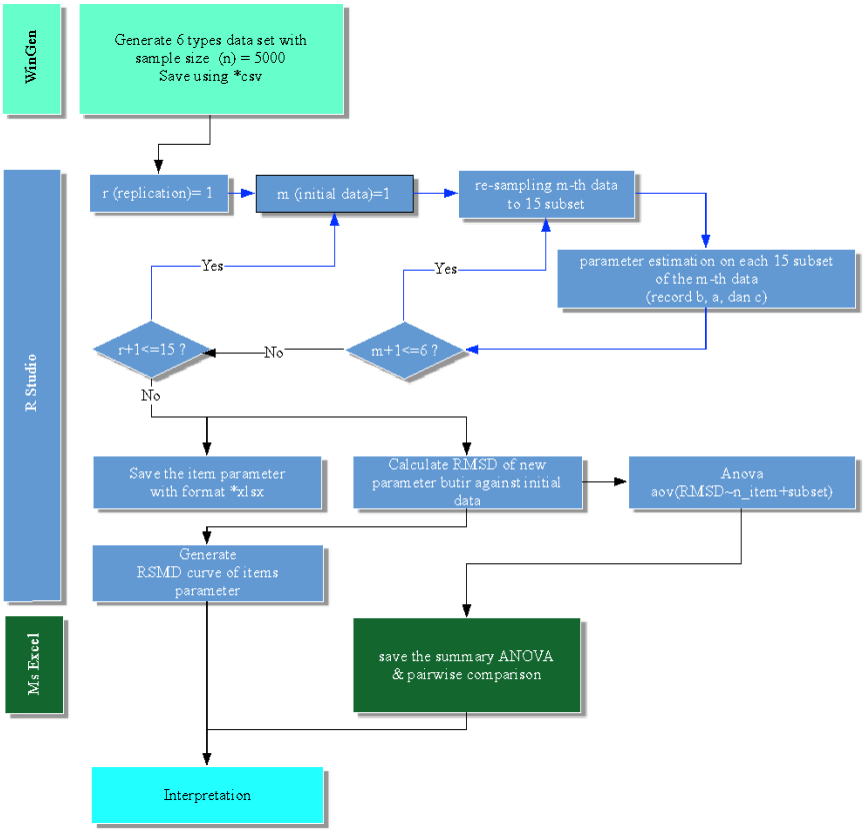
Research Objective

Studies on sample size in IRT have shown the need for further exploration in determining sample size for calibrating IRT test items, impacting parameter estimation and the assumption of item parameter invariance. Unlike previous studies, which generated simulation data separately and employed various IRT models for parameter estimation, our study utilizes six datasets already fitting the 3PL IRT model, each with a sample size 5000. The primary goal of this study is to evaluate the accuracy of item parameters estimation when employing the 3PL IRT model. We address three key research questions: (1) does sample size and test length affect item parameter estimation accuracy in the IRT 3-PL model?; (2) what test length is necessary to uphold item parameter accuracy in the IRT 3-PL model?; and (3) what is the required sample size to preserve item parameter accuracy in the IRT 3-PL model?

Methods

In this simulation study, we employed the 'mirt' package in RStudio (Chalmers, 2012; Djidu et al., 2022; Fernández-Ballesteros, 2012) to analyze data generated using WinGen. We estimated item parameters for six datasets with common sample sizes (5000 responses) but varying test lengths (10, 15, 20, 25, 30, and 40 items).

For each dataset, we performed resampling by randomly selecting response subsets. Subsequently, we iteratively estimated item parameters using the IRT 3-PL model on all these subsets. We then recorded and compared the resulting item parameters with the initial ones. The simulation study's stages are illustrated in Figure 1.



Source: Personal data (2023)

Figure 1. Flowchart Item Parameter Estimation in Simulation Studies

In our simulation, we employed three software tools: WinGen (Han, 2007; Han, 2007) to generate simulation data with consistent parameters for each dataset. The only variation among the initial six datasets was in test length, ranging from 10 to 40. RStudio was then utilized for random resampling of 5000 samples from the initial data, item parameter estimation, quantifying item parameter bias, and performing ANOVA tests to assess the impact of test length and sample size on item parameter deviations from their original values. Additionally, MS Excel was used to store the analysis results, including the summary of ANOVA and pairwise comparisons.

Data Generation

The simulation commenced by creating six datasets with specific characteristics using WinGen (Han, 2007; Han, 2007). Each dataset shared common attributes: (1) number of examinees = 5000, (2) distribution = normal, (3) mean = 0, (4) standard deviation = 1, (5) number of response categories = 2, and (6) model = 3PLM (3-Parameter Logistic Model). The three item parameters (par.a, par.b, and par.c) were assigned particular distributions and mean values: (1) par.a: normal distribution with a mean of 0.85 and a standard deviation of 0.1, (2) par.b: normal distribution with a mean of 0.65 and a standard deviation of 0.6, and (3) par.c: uniform distribution with a minimum value of 0.001 and a standard deviation of 0.05. Test length varied across the six dataset types, resulting in test lengths of 10 items, 15 items, 20 items, 25 items, 30 items, and 40 items, respectively. The generated data consisted of responses in a dichotomous format (1/0), representing correct/incorrect answers.

Resampling the Initial Data

The initial dataset included 5000 samples. Each of the six initial datasets underwent resampling across 15 scenarios, resulting in 90 new data subsets. The sample sizes for these subsets were systematically decreased from 5000 to as low as 200. The subsets were created with reduced sample sizes: 4500, 4000, 3500, 3000, 2500, 2000, 1500, 1000, 800, 600, 500, 400, 300, 250, and 200. This reduction in sample sizes allowed for a comprehensive data exploration under various conditions during the resampling process.

The resampling process was conducted randomly using the "sample_n(data, n-sample target)" function from the 'dplyr' package in RStudio (Wickham et al., 2022). Each subset (resampling) was replicated 15 times, resulting in a total of 2025 data subsets. For each item, the parameters were estimated 225 times (15×15). In summary, this resampling procedure generated 2025 data subsets, with item parameters estimated 225 times through this iterative process.

Estimating Items Parameters

The initial data is used for estimating item parameters (b, a, and c) using the IRT 3-PL model. These parameters are saved as reference parameters (b0, a0, and c0) and serve as a basis for evaluating item parameters obtained from resampled subsets of data. The estimation of item parameters in both the initial data and data subsets is conducted using the 'mirt' package in RStudio (Fernández-Ballesteros, 2012).

For each of the six test lengths (10, 15, 20, 25, 30, and 40 items), item parameters (b, a, and c) are estimated in the initial data. This results in a set of estimates labelled accordingly (e.g., 10.b0, 10.a0, 10.c0 for a test length of 10 items). In each test length scenario, item parameters are estimated across 15 resampled data subsets, with each subset being replicated 15 times. This leads to 225 dataset estimates for each item (b, a, and c) at a given test length.

In summary, this process allows for a comprehensive analysis of the accuracy and precision of item parameters across various test lengths and sample sizes, ensuring a thorough evaluation of the IRT model's performance.

Evaluation of Item Parameter Estimation Accuracy

The evaluation of item parameter estimation accuracy in this study aligns with the approach introduced by Paek et al. (2021). They employed the Root Mean Squared Difference (RMSD) and Mean Absolute Difference (MAD) to gauge the precision and bias of the IRT 2-PL item parameters derived from re-sampled data. Notably, the simulation results by Paek et al. (2021) revealed that the resulting RMSD and MAD exhibited no significant differences. This finding implies that due to their comparable performance, either RMSD or MAD can be chosen to evaluate the accuracy of item parameter estimation in the context of IRT. This method is akin to that used by Wells et al. (2002), who utilized RMSD to assess the accuracy of ability estimation.

In this study, RMSD is employed as the method of choice to evaluate the accuracy of item parameters in the simulation, ensuring consistent evaluation criteria are applied throughout the analysis. For each item, the RMSD of the parameter estimation results on the each item is calculated using the following equation.

$$RSMD(i) = \sqrt{\frac{\sum_{r=1}^{15}\left(\hat{\xi}_{ir} - \hat{\xi}_{i0}\right)^2}{R}}$$

Where "i" represents the item number, "r" represents the replication (1, 2, 3, …, R), ξ_{ir} is the estimated item parameter result for item i in replication r, and ξ_{i0} is the initial (reference) parameter for item i. RMSD(i) values are computed for each item on 15 subsets of the initial data across all six test lengths. This process generates 90 RMSD(i) values, one for each subset. The entire calculation procedure is carried out within RStudio.

RMSD serves as a dual-purpose metric, indicating both the accuracy of parameter estimation and the extent of parameter bias introduced by re-sampling the initial parameters. A higher RMSD value signifies a significant deviation of parameter magnitude from the initial values, indicating lower accuracy in the estimation results. Conversely, a lower RMSD value suggests minimal bias and higher accuracy in parameter estimation.

To evaluate the effect of sample sizes and test length on item parameter accuracy, we conducted a one-way analysis of variance (ANOVA). This analysis scrutinizes each factor individually, considering the sample size of the re-sampled data (subset) and the test length (n-items). The ANOVA uses the 'aov' function within the 'stats' package in RStudio.

If the two-factor ANOVA reveals significant differences in RMSD due to the test length and/or sample size factors, a post hoc analysis, specifically Tukey's Honestly Significant Difference (HSD) test, is employed. Tukey's HSD is a commonly utilized method for comparing all pairwise means following the detection of significant differences in ANOVA results. Through the application of Tukey's HSD, we can pinpoint which combinations of sample sizes or test lengths exhibit significant deviations in the RMSD of item parameters compared to the initial data. This approach provides in-depth insights into the specific distinctions among these groups following the initial ANOVA analysis, which has identified significant differences.

The evaluation of parameter estimation accuracy depends on the significance of the ANOVA results, which are used to evaluate RMSD variance. The null hypothesis under examination posits that the RMSD of item parameters for all sample sizes is zero. If the analysis produces significant results (p-value < 0.05), it indicates a noteworthy RMSD in item parameters for a specific sample size compared to the initial data.

Results and Discussion

Results

The initial analysis aimed to investigate the effect of sample sizes and test length on item parameter accuracy, addressing whether they affect item parameter estimation accuracy in the IRT 3-PL model. The data analysis in Table 1 shows that sample sizes significantly affect the RMSD of item parameters, including item difficulty (b), item discrimination (a), and pseudo-guessing (c). Additionally, the test length also significantly affects the RMSD of all item parameters.

Further analysis, as displayed in Figure 2, Figure 3, and Figure 4 demonstrates an increasing trend in RMSD values. This increasing trend indicates that both sample size and test length have a significant impact on the decrease in accuracy in estimating parameters b, a, and c. These findings demonstrate that both sample size and test length are crucial factors influencing the accuracy of item parameter estimation in the 3-PL IRT model. This indicates that the sample sizes used for test calibration and test length play significant roles in this accuracy.

Table 1. ANOVA of RSMD (by: Sample Sizes & Test Length)

Factor	Parameter	Df	Mean Sq	F value	Pr(>F)
Sample sizes					
	b	14	0.039	33.247	0.000
	a	14	0.551	24.811	0.000
	c	14	0.003	107.088	0.000
Test length					
	b	5	0.013	11.324	0.000
	a	5	0.095	4.263	0.002
	c	5	0.000	17.626	0.000

Source: Personal data (2023)

Considering the significant differences in RMSD item parameters across various sample sizes and test lengths, we conducted further pairwise comparisons (Table 4 and Table 5) to identify distinct groups and evaluate the minimum sample size required to maintain item parameter accuracy. Additionally, pairwise comparison was employed to explore the accuracy of item parameters to the test length. This pairwise analysis simultaneously addresses the second and third questions in this study, which are related to "What test length is necessary to uphold item parameter accuracy in the IRT 3-PL model?" and "What is the required sample size to preserve item parameter accuracy in the IRT 3-PL model?".

Table 2. Pairwise RMSD Comparison (by: Test Length)

Test Length (n)	10 Item	15 Item	20 Item	25 Item	30 Item
Param b					
10 item	-				
15 item	0.003	-			
20 item	0.004	1.000	-		
25 item	0.001	1.000	0.999	-	
30 item	0.951	0.000	0.000	0.000	-
40 item	0.000	0.978	0.960	0.998	0.000
Test length (n)	**10 item**	**15 item**	**20 item**	**25 item**	**30 item**
Param a					
10 item	-				
15 item	0.183	-			
20 item	0.012	0.889	-		
25 item	0.013	0.899	1.000	-	
30 item	0.478	0.993	0.569	0.587	-

http://journal.uinjkt.ac.id/index.php/jp3i

Test Length (n)	10 Item	15 Item	20 Item	25 Item	30 Item
40 item	0.002	0.597	0.995	0.993	0.261
Test length (n)	**10 item**	**15 item**	**20 item**	**25 item**	**30 item**
Param c					
10 item	-				
15 item	0.000	-			
20 item	0.000	0.012	-		
25 item	0.000	0.060	0.992	-	
30 item	0.000	0.935	0.139	0.400	-
40 item	0.000	0.994	0.054	0.199	0.999

Source: personal data (2023)

Table 2 displays pairwise comparisons of RMSD between data groups categorized by test length factors. A significant p-value indicates significant variations in RMSD among these groups. Specifically, the comparison between test lengths 10 and 30 reveals that the RMSD parameter b is not significantly different (p-value > 0.05). However, it significantly differs from the other four test lengths. Moreover, pairwise comparisons for all combinations of test lengths (15, 20, 25, and 40) exhibit p-values > 0.05, indicating that RMSD parameter b for these four test lengths is not significantly different. Pairwise analysis based on item discrimination (a) reveals significant differences in RMSD between test length 10 and test lengths 20, 25, and 40. Meanwhile, pairwise comparisons for all combinations of test lengths (15, 20, 25, 30, and 40) exhibit p-values > 0.05, indicating that RMSD parameter a for these five test lengths is not significantly different. Furthermore, pairwise analysis based on pseudo-guess (c) yields results similar to the previous two parameters. RMSD for the test length 10 indicates a significant difference from the other five test lengths.

Based on these results, the most significant RMSD differences were observed in data with a test length of 10. These differences also tend to lead to estimation inaccuracies, as the RMSD values for data with a test length of 10 significantly increase with decreasing sample size (see Figure 2, Figure 3, and Figure 4). These findings imply that a minimum of 15 test items is required to obtain unbiased estimates of item parameters.

Table 3. Pairwise RMSD Comparison (by: Sample Sizes)

Sample sizes	4500	4000	3500	3000	2500	2000	1500	1000	800	600	500	400	300	250
param b	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**-**
4000	1.00	-	-	-	-	-	-	-	-	-	-	-	-	-
3500	0.98	1.00	-	-	-	-	-	-	-	-	-	-	-	-
3000	0.85	1.00	1.00	-	-	-	-	-	-	-	-	-	-	-
2500	0.41	0.95	1.00	1.00	-	-	-	-	-	-	-	-	-	-
2000	0.13	0.67	0.93	1.00	1.00	-	-	-	-	-	-	-	-	-
1500	0.04	0.36	0.71	0.93	1.00	1.00	-	-	-	-	-	-	-	-
1000	0.00	0.08	0.26	0.53	0.92	1.00	1.00	-	-	-	-	-	-	-
800	0.00	0.00	0.01	0.03	0.19	0.51	0.81	0.99	-	-	-	-	-	-
600	0.00	0.00	0.00	0.00	0.01	0.05	0.16	0.54	1.00	-	-	-	-	-
500	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.12	0.85	1.00	-	-	-	-
400	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.12	0.73	0.99	-	-	-
300	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.31	0.98	-	-
250	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.09	0.87	-
200	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.41	1.00
Sample sizes	**4500**	**4000**	**3500**	**3000**	**2500**	**2000**	**1500**	**1000**	**800**	**600**	**500**	**400**	**300**	**250**
param a	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**-**
4000	1.00	'-	-	-	-	-	-	-	-	-	-	-	-	-
3500	1.00	1.00	'-	-	-	-	-	-	-	-	-	-	-	-
3000	1.00	1.00	1.00	'-	-	-	-	-	-	-	-	-	-	-

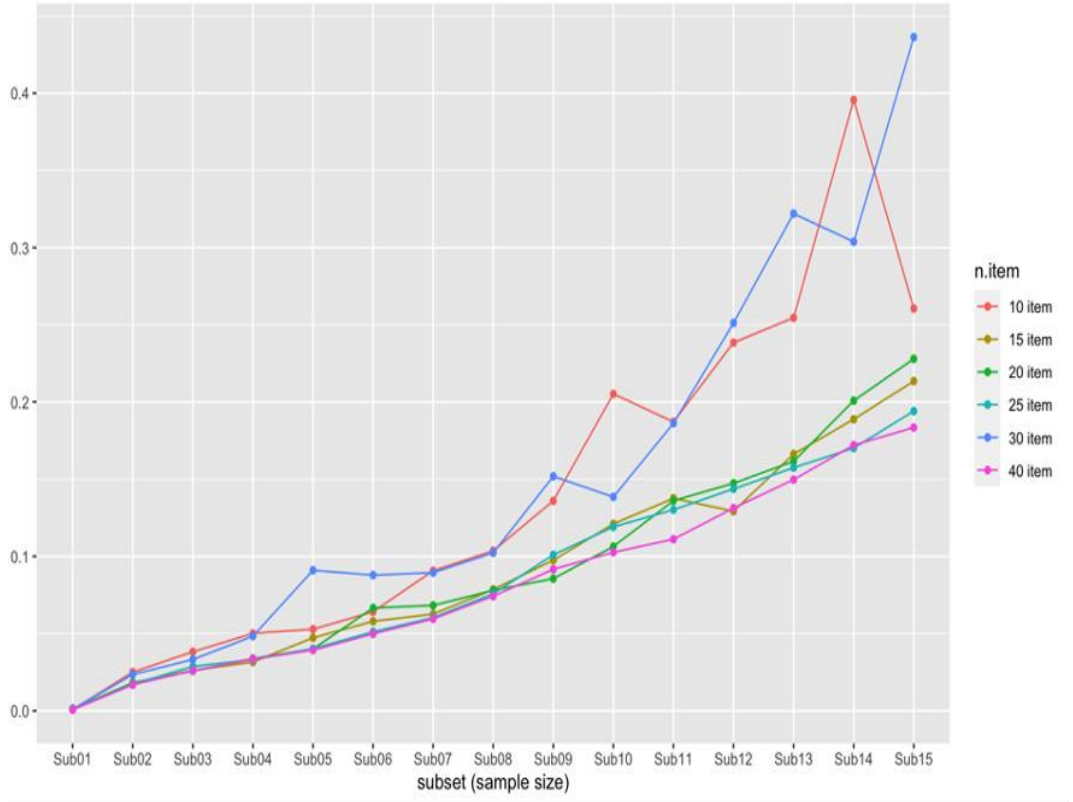
Sample sizes	4500	4000	3500	3000	2500	2000	1500	1000	800	600	500	400	300	250
2500	1.00	1.00	1.00	1.00	'-	-	-	-	-	-	-	-	-	-
2000	1.00	1.00	1.00	1.00	1.00	'-	-	-	-	-	-	-	-	-
1500	1.00	1.00	1.00	1.00	1.00	1.00	'-	-	-	-	-	-	-	-
1000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	'-	-	-	-	-	-	-
800	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	'-	-	-	-	-	-
600	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	'-	-	-	-	-
500	0.91	0.96	0.98	0.99	0.99	1.00	1.00	1.00	1.00	1.00	'-	-	-	-
400	0.87	0.94	0.96	0.98	0.99	1.00	1.00	1.00	1.00	1.00	1.00	'-	-	-
300	0.27	0.38	0.43	0.51	0.57	0.68	0.76	0.82	0.94	0.99	1.00	1.00	'-	-
250	0.01	0.02	0.02	0.03	0.04	0.06	0.08	0.11	0.23	0.40	0.65	0.72	1.00	'-
200	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.03	0.04	0.36	0.98
Sample sizes	**4500**	**4000**	**3500**	**3000**	**2500**	**2000**	**1500**	**1000**	**800**	**600**	**500**	**400**	**300**	**250**
param c	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**-**
4000	0.74	-	-	-	-	-	-	-	-	-	-	-	-	-
3500	0.09	1.00	-	-	-	-	-	-	-	-	-	-	-	-
3000	0.00	0.59	1.00	-	-	-	-	-	-	-	-	-	-	-
2500	0.00	0.18	0.89	1.00	-	-	-	-	-	-	-	-	-	-
2000	0.00	0.00	0.12	0.74	0.99	-	-	-	-	-	-	-	-	-
1500	0.00	0.00	0.00	0.03	0.16	0.93	-	-	-	-	-	-	-	-
1000	0.00	0.00	0.00	0.00	0.00	0.07	0.92	-	-	-	-	-	-	-
800	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.64	-	-	-	-	-	-
600	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.96	-	-	-	-	-
500	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.76	-	-	-	-
400	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.47	1.00	-	-	-
300	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.24	0.49	-	-
250	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.14	-
200	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.90

Source: Personal data (2023)

The analysis of pairwise comparisons for RMSD item parameters based on sample sizes (Table 3) underscores the remarkable sensitivity of the IRT 3-PL to variations in sample sizes. Specifically, significant differences in item parameters emerge at sample sizes of 3000 (pseudo guessing), 1500 (difficulty parameter), and 250 (item discrimination). It is imperative to interpret these RMSD comparison results holistically since any significant difference in any of the parameters compared to the item parameters in the initial data signifies a substantial overall deviation in the item parameters.

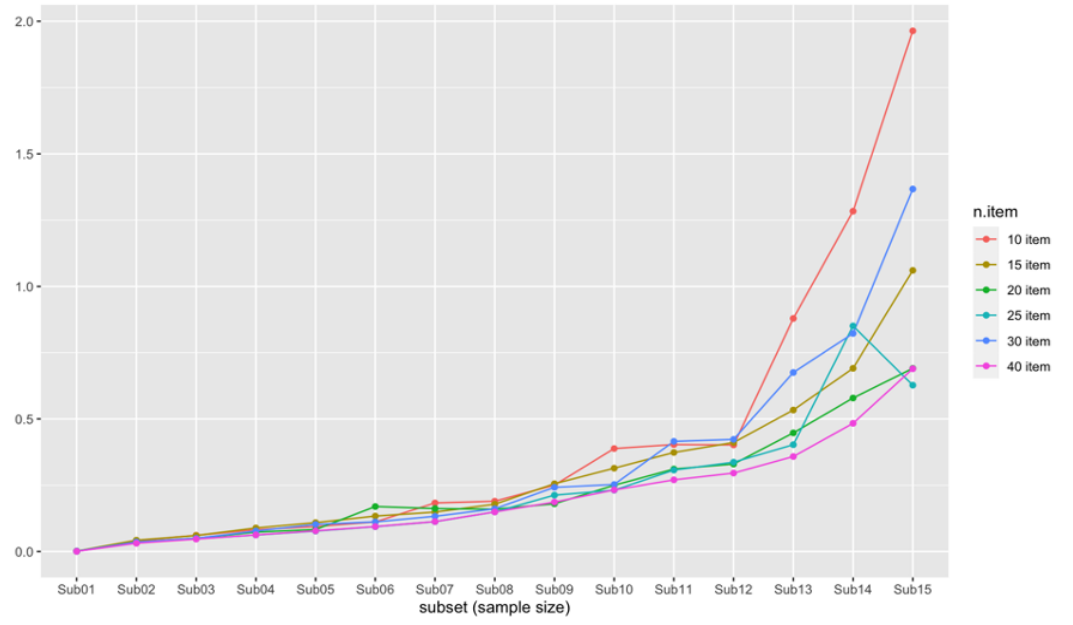
Figure 2 visually represents the variation in RMSD of item parameters from the initial data to the 15th subset (sample size). As the sample size decreases, the RMSD increases, indicating a growing bias between parameters b in smaller sample size compared to the initial b. ANOVA reveals a significant change in the value of b at the seventh subset, corresponding to a sample size of 1500 or smaller. However, the estimation results in the first six subsets (4500, 4000, 3500, 3000, 2500, and 2000) consistently yield similar value. This suggests that item difficulty stabilizes in the 3-PL IRT when the sample size is a minimum of 2000. Additionally, Figure 2 highlights the influence of test length on the RMSD of item parameters in the data. Test lengths of 10 and 30 items exhibit the most significant changes compared to other lengths, aligning with the findings from the pairwise comparison analysis presented in Table 4.

http://journal.uinjkt.ac.id/index.php/jp3i



Source: personal data (2023)

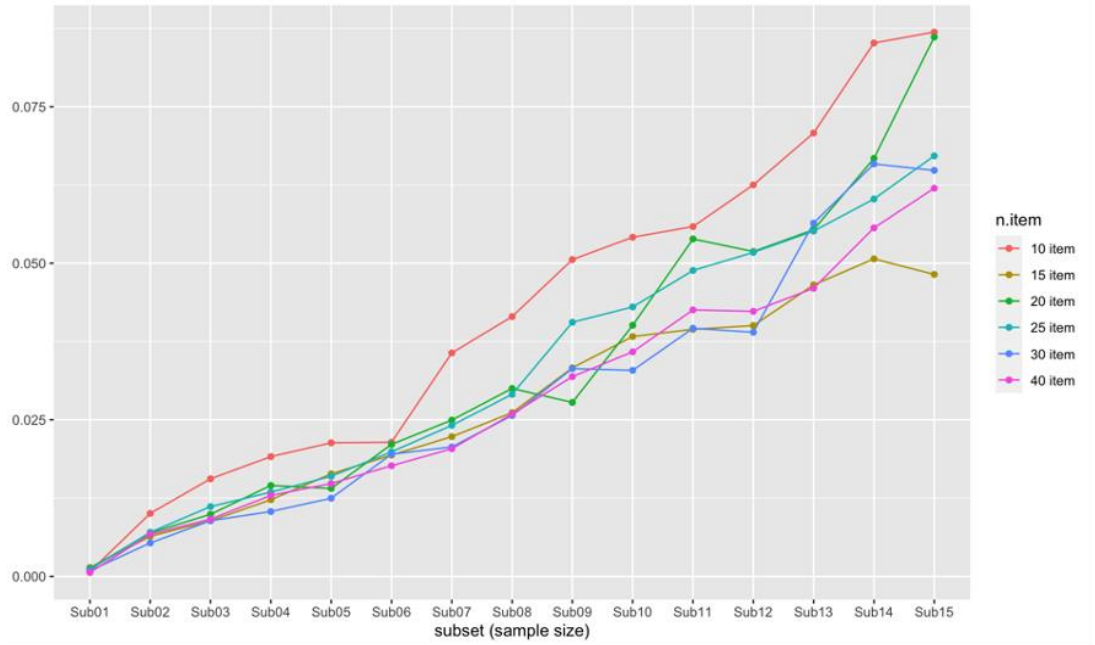
Figure 2. RMSD Parameter b for 15 Sample Sizes with 225 Replications



Source: personal data (2023)

Figure 3. RMSD Parameter a for 15 Sample Sizes with 225 Replications

Figure 3 demonstrates the stability of parameter a compared to the other two parameters (b and c) in subsets 01 to 13. The RMSD of parameter a remains relatively stable from data with test length=10 up to test length=15, staying consistent until subset 12 (sample size=400). However, data with test length=10 and test length=30 exhibit higher RMSD values than the other two parameters.



Source: Personal data (2023)

Figure 4. RMSD Parameter c for 15 Sample sizes with 225 Replications

Figure 4 illustrates the RMSD of parameter c, similar to Figure 2 and Figure 3. Examining the RMSD of parameter c, it becomes apparent that data with test length = 10 consistently exhibits the highest RMSD among all test lengths. Conversely, data with test length = 40 consistently displays the lowest RMSD. Figure 4 portrays the magnitude of the change in RMSD from subset 3 to subset 15. Unlike parameters b and a, where the RMSD change undergoes significant variations in smaller sample sizes, the RMSD of parameter c exhibits relatively consistent changes across different sample sizes.

Discussions

The simulation results highlight the significant influence of test length on the accuracy of item parameters in the IRT 3-PL model, a critical consideration given its impact on participants' psychological states (Ackerman & Kanfer, 2009). Consequently, careful test length selection is essential for ensuring precise item parameter estimation in pilot studies.

Notably, the study's simulation results reveal that a test with a length of 40 provides the most accurate item parameters. Conversely, shorter tests result in unstable parameter estimation. However, these findings do not mandate lengthy tests for measuring latent attributes; instead, they offer insights into the ideal number of items for pilot study calibration. Previous research (Ackerman & Kanfer, 2009; Şahin & Anıl, 2017; Wells et al., 2002) has also explored the impact of test length on ability estimation, with some studies finding no significant effect.

Sample size emerges as the most influential factor affecting item parameter stability in the IRT 3-PL model. Parameters b, a, and c remain stable when the sample size surpasses 3000, based on estimations from 225 replications for each item. Consequently, large sample sizes are imperative for accurately calibrating item parameters in the IRT 3-PL model. Small sample sizes in pilot studies will likely yield item parameter estimates with significant bias.

It is important to note that this study did not explore specific attributes within the sample data, such as distribution or other characteristics that may impact parameter estimation accuracy. The study employed random sampling on a subset of the data. Therefore, further research is needed to investigate how different data attributes affect the stability of item parameters in the IRT 3-PL model.

http://journal.uinjkt.ac.id/index.php/jp3i

The findings of this study are closely aligned with those of Paek et al. (2021), who examined item parameters in IRT 2-PL and Rasch models. Paek et al. (2021) also observed that the difficulty parameter (b) was sensitive to sample size in IRT models with item discrimination parameters (a). This sensitivity to sample size was similarly observed in our study, with the difficulty parameter exhibiting a significant difference in the sixth sample size, while the item discrimination parameters (a) showed significant differences for the 13th sample size.

Although the RMSD characteristics in this simulation exhibit similarities with Şahin and Anıl (2017), the conclusions about the accuracy of estimating item parameters using the 3-PL model differ between the two studies. In our study, the RMSD value of item parameters can exceed 0.33 after the 12th subset, corresponding to a sample size of 350, while Şahin and Anıl (2017) concluded that a minimum sample size of 350 and test lengths of 30 resulted in precise parameter estimation with RMSD<0.33.

In our study, different criteria were used, employing ANOVA to assess whether RMSD obtained from subsets with large samples could still be maintained. Consequently, to achieve precise parameter estimation, the minimum sample sizes identified in our study were larger than those reported by Şahin and Anıl (2017).

Based on the stability analysis of item parameters (b, a, and c) in this simulation, two crucial conclusions can be drawn to obtain unbiased item parameters. Firstly, the minimum sample size for 3-PL IRT models required for pilot study/calibration should be at least 3000. Larger sample sizes enhance accuracy and stability in parameter estimation, reducing bias. Secondly, for pilot study/calibration, it is advisable to use a test length of at least ten items, with test lengths of 25 or 40 being preferred.

The results underscore the significance of achieving the assumption of invariance of item parameters (Retnawati, 2014; Stenbeck et al., 1992), which necessitates conducting a pilot study with large sample sizes. Small sample sizes can lead to imprecise item parameter estimates, affecting the accuracy of measuring latent attributes. Therefore, substantial sample sizes are crucial for reliable and accurate item parameter estimation, ensuring a precise assessment of latent attributes.

Furthermore, the quality of the items and measuring instruments is intricately linked to the skills and expertise of those involved in item development. While this study offers a quantitative approach to assessing item attributes through statistical testing, it is vital to acknowledge that the procedure for evaluating item content quality plays a central role in determining overall instrument quality. Thus, adhering to good item development guidelines is highly recommended to produce high-quality instrument items.

The findings from this simulation study provide valuable guidance for planning large-scale pilot studies for calibration purposes. By applying the insights gained from this research, practitioners and researchers can enhance the effectiveness and accuracy of the measurement process, ultimately leading to more robust and reliable results in the field of psychometrics.

Conclusion

The study's outcomes yield three main conclusions regarding sample size, test length, and sensitivity to sample size. Firstly, both sample size and test length significantly impact the accuracy of item parameters in the IRT 3-PL model, with larger sample sizes and test lengths leading to more stable parameter estimations. Secondly, a limited number of items in a short test results in unstable and biased parameter estimates. To enhance accuracy and reliability, it is advisable to use a test length of 25 or 40 items during pilot studies. Thirdly, estimating item parameters using the IRT 3-PL model is highly sensitive to sample sizes, with smaller sample sizes introducing greater bias in parameter estimates. Achieving precise calibration necessitates a minimum sample size of 3000 for estimating parameters (b,

a, and c) in the 3-PL IRT model. By incorporating these conclusions, test developers can improve the quality and accuracy of their measurement instruments. Diligence in considering sample size, test length, and calibration procedures enhances the reliability of psychometric evaluations, offering valuable insights into participants' latent attributes.

Limitations and Suggestions for Further Research

The simulations in this study involved 225 replications for each item's parameter estimation, spanning six test length scenarios and 15 different sample sizes. The results from these replications offer insights into the required sample sizes for precise item parameter estimation. However, it's important to acknowledge that the simulation data used in this study followed the same distribution characteristics, specifically the normal distribution. This limitation should be recognized, and further studies using simulated data with diverse characteristics are needed for a more comprehensive and robust understanding.

Future research should focus on determining the minimum sample size and considering participant characteristics in pilot studies. Understanding how participant characteristics impact parameter estimation is crucial for improving the accuracy of measuring instruments. Regarding the sensitivity of parameter estimation results caused by test length differences, this study did not investigate the instability observed at a test length of 30. It would be intriguing to explore the reasons behind this instability. Conducting further empirical studies to identify the underlying causes is a potential avenue for research.

In summary, this study provides valuable insights into the effects of sample size and test length on item parameter estimation while also pointing out areas for future research and enhancement. Addressing these limitations and conducting additional studies will contribute to a better understanding and more accurate item parameter estimation in the context of the IRT 3-PL model.

Acknowledgment

We would like to express the highest gratitude to Higher Education Funding Center, The Ministry of Education, Culture, Research, and Technology (Balai Pembiayaan Pendidikan Tinggi (BPPT)-Kemdikbudristek) of the Republic of Indonesia, and The Indonesia Endowment Funds for Education (LPDP) for providing the financial support (Indonesia Educational Scholarships) to the first author to pursue doctoral education and complete this research.

Conflict of Interest

The authors declare no conflict of interest.

Authors Contribution

Conceptualization, HD; methodology, HD, HR, and H; software, HD; validation, HR, and H; formal analysis, HD and HR; investigation, HD; resources, HD; data curation, HD; writing—original draft preparation, HD; writing—review and editing, HD; visualization, HD; supervision, HR and H; project administration, HD. All authors have read and agreed to the published version of the manuscript.

References

Ackerman, P. L., & Kanfer, R. (2009). Test length and cognitive fatigue: An empirical examination of effects on performance and test-taker reactions. *Journal of Experimental Psychology: Applied*, *15*(2), 163–181. https://doi.org/10.1037/a0015719

Baker, F. B. (1985). Book review: Item response theory: Principles and applications. In *Applied*

Psychological Measurement (Vol. 9, Issue 3). Nijhof'f Publishing. https://doi.org/10.1177/014662168500900315

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

Divgi, D. R. (1984). Does small N justify use of the Rasch model. *Annual Meeting of the American Educational Research Association, New Orleans*.

Djidu, H., Ismail, R., Sumin, Rachmaningtyas, N. A., Imawan, O. R., Suharyono, Aviory, K., Prihono, E. W., Kurniawan, D. D., Syahbrudin, J., Nurdin, Marinding, Y., Firmansyah, Hadi, S., & Retnawati, H. (2022). *Analisis instrumen penelitian dengan pendekatan teori tes klasik dan modern menggunakan program R*.

Fernández-Ballesteros, R. (2012). Multidimensional item response theory. In *Encyclopedia of Psychological Assessment*. https://doi.org/10.4135/9780857025753.n128

Feuerstahler, L. M. (2022). Metric stability in item response models. *Multivariate Behavioral Research*, *57*(1), 94–111. https://doi.org/10.1080/00273171.2020.1809980

Han, K. T. (2007a). WinGen. In *Computer software]. Amherst, MA: University of Massachusetts at Amherst*.

Han, K. T. (2007b). WinGen: Windows software that generates item response theory parameters and item responses. *Applied Psychological Measurement*, *31*(5), 457–459. https://doi.org/10.1177/0146621607299271

Paek, I., Liang, X., & Lin, Z. (2021). Regarding item parameter invariance for the Rasch and the 2-parameter logistic models: An investigation under finite non-representative sample calibrations. *Measurement*, *19*(1), 39–54. https://doi.org/10.1080/15366367.2020.1754703

Retnawati, H. (2014). *Teori respon butir dan penerapannya untuk peneliti, praktis pengukuran dan penguji*. Parama Publishing. http://staff.uny.ac.id/sites/default/files/pendidikan/heri-retnawati-dr/teori-respons-butir-dan-penerapanya-135hal.pdf

Şahin, A., & Anıl, D. (2017). The effects of test length and sample size on item parameters in item response theory. *Kuram ve Uygulamada Egitim Bilimleri*, *17*(1), 321–335. https://doi.org/10.12738/estp.2017.1.0270

Stenbeck, M., Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1992). Fundamentals of item response theory. In *Contemporary Sociology* (Vol. 21, Issue 2). SAGE Publications. https://doi.org/10.2307/2075521

Wainer, H., & Wright, B. D. (1980). Robust estimation of ability in the Rasch model. *Psychometrika*, *45*(3), 373–391. https://doi.org/10.1007/BF02293910

Wells, C. S., Subkoviak, M. J., & Serlin, R. C. (2002). The effect of item parameter drift on examinee ability estimates. *Applied Psychological Measurement*, *26*(1), 77–87. https://doi.org/10.1177/0146621602261005

Wickman, H., François, R., Henry, L., & Muller, K. (2021). dlpyr: A grammar of data manipulation. In *CRAN Repository* (Vol. 3, pp. 1–2). https://cran.r-project.org/package=dplyr

Wise, S. L. (1991). The utility of a modified one-parameter IRT model with small samples. *Applied Measurement in Education*, *4*(2), 143–157. https://doi.org/10.1207/s15324818ame0402_4

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, *5*(2), 245–262. https://doi.org/10.1177/014662168100500212

Appendix

R Code

R code to run the analysis available at: https://bit.ly/HD-RStudioCodeSimulation

Data

Data available at: https://bit.ly/HD-DataSimulation

http://journal.uinjkt.ac.id/index.php/jp3i