Tabaran Institute of Higher Education
International Journal of Language Testing
Vol. 15, No. 2; 2025, October

ISSN 2476-5880
https://www.ijlt.ir

Three-Parameter Item Response Theory Analysis of the Multiple-Choice Items in PIRLS 2016

Mulyono[1]*, Rahma Widyana[2], Puspa Mirani Kadir[3], Wulida Wahidatul Masruria[4], M. Baihaqi[5], Engeline Chelsya Setiawan[6]

ARTICLE INFO

Article History:
Received: February 2025
Accepted: March 2025

KEYWORDS
3PL IRT
Guessing
ICC
IRT
PIRLS

ABSTRACT

Progress in International Reading Literacy Study (PIRLS) is an international assessment that measures the reading literacy of fourth-grade students (aged 9-10 years old). PIRLS aims to evaluate and compare the reading abilities of students across different countries. It assesses how well students can understand and interpret written texts, which is fundamental to their overall educational development. In this study, psychometric analyses were run on a portion of the multiple-choice items of PIRLS 2016 taken by 4th graders in the USA. The 3PL item response theory model was utilized to examine the test. Discrimination, difficulty, and guessing parameters were estimated along with the fit values, reliability, item characteristic curves, and item-person map. M2, CFI, TLI, and RMSE statistics showed that the test is reliable and the model, overall, fits the data. Item fit statistics outfit and infit showed that most of the items fit the 3PL model. Findings showed that while all the items have acceptable discrimination values, two items have unacceptable guessing parameters. Examination of the ICCs showed that graphical displays are important, in addition to numerical values, for examining item quality. Item-person map showed that items do not target the whole ability scale.

1. Introduction

The Progress in International Reading Literacy Study (PIRLS) is a global assessment and research initiative aimed at evaluating reading proficiency among fourth-grade students,

[1] State University of Surabaya, Indonesia. https://orcid.org/ 0000000316116617. Email: mulyono42036925@gmail.com
[2] Universitas Mercu Buana Yogyakarta, Indonesia. https://orcid.org/0000-0002-3807-3470
[3] Universitas Padjadjaran, Indonesia. https://orcid.org/0000-0002-6466-5189
[4] Boston University, USA. https://orcid.org/0009-0009-5768-4347
[5] UIN Sunan Ampel Surabaya, Indonesia. https://orcid.org/0009-0007-7171-8449
[6] Department of Health Policy and Administration, Faculty of Public Health Universitas Airlangga, Surabaya 60115, Indonesia. https://orcid.org/0009-0008-8985-442X

Cite this paper as: Mulyono, M., Widyana, R., Mirani Kadir, P., Wahidatul Masruria, W., Baihaqi, M., & Chelsya Setiawan, E. (2025). Three-Parameter Item Response Theory Analysis of the multiple-choice items in PIRLS 2016. *International Journal of Language Testing*, *15*(2), 109–119.
https://doi.org/10.22034/ijlt.2025.502694.1414

along with examining school and teacher practices related to instruction. Conducted every five years since 2001, PIRLS involves fourth-grade students completing a reading test and a questionnaire that explores their attitudes towards reading and their reading behaviors. Additionally, teachers and school principals are surveyed to collect information about students' school experiences in developing reading literacy. PIRLS offers valuable benchmark data that allows countries to compare their students' performance with that of students worldwide. It enables educators, researchers, and policymakers to explore educational practices in other systems that could be applied locally, contributing to ongoing efforts to enhance the quality of education for all students (Mullis & Martin, 2015).

PIRLS 2016 is the fourth assessment in the ongoing series, succeeding those held in 2001, 2006, and 2011. Sixty-one entities, comprising 50 countries, and 11 regions took part in PIRLS 2016. For countries that have participated in previous assessments since 2001, PIRLS 2016 provides an opportunity to track changes in reading achievement over the years: 2001, 2006, 2011, and 2016 (Mullis et al., 2017).

The PIRLS 2016 assessment is guided by the PIRLS 2016 Assessment Framework (Mullis & Martin, 2015), which was developed with input from the participating countries. This framework focuses on two primary reading goals: engaging with literary texts and acquiring and using information. It identifies four key reading comprehension processes: locating and retrieving explicitly stated information, making straightforward inferences, interpreting and integrating ideas and information, and evaluating and critiquing content and text elements. For PIRLS 2016, a nationally representative sample of about 4,000 students from 150 to 200 schools from each participating country was assessed. Overall, the assessment involved approximately 319,000 students, 310,000 parents, 16,000 teachers, and 12,000 schools across all the countries (Mullis et al., 2017).

1.1 Item Types in PIRLS

Students' reading comprehension is evaluated through questions that follow each text, using two item formats, namely, multiple-choice (MC) and constructed-response questions. Each MC question is worth one point, while constructed-response questions are valued at one, two, or three points, depending on the level of understanding required. In PIRLS, MC questions can account for up to half of the total points available. The choice between MC and constructed-response questions is based on the comprehension process being assessed and which format best allows students to show their understanding.

MC questions present four answer options, only one of which is correct. These questions can assess any aspect of comprehension, but they are less effective for evaluating students' ability to provide detailed explanations or complex interpretations. For fourth graders, questions are crafted to be developmentally appropriate, with clear and concise wording. Response options are also brief to reduce reading difficulty, and incorrect answers are designed to be plausible but not misleading (Mullis & Martin, 2015).

One of the disadvantages of MC items is that they can be guessed. In this study, the guessing problem in the MC question of PIRLS 2016 was examined in a small portion of the test using the 3-parameter logistic item response theory model (3PL IRT, Birnbaum, 1968) with the USA data. The 3-parameter IRT model is particularly useful in situations where multiple-choice items are used, and guessing can be a factor. It provides a more nuanced understanding of item performance and individual ability compared to simpler models like the 1-parameter (Rasch) or 2-parameter IRT models, which do not account for guessing.

2. The 3PL IRT Model

The 3-PL IRT model (Birnbaum, 1968) is a psychometric model used to analyze the relationship between an individual's latent trait (often referred to as ability or proficiency) and their performance on test items. This model is commonly applied in educational testing, psychological assessments, and other fields where the measurement of latent traits is important. The item response function for the 3PL model is:

$$P(X_i = 1|\theta) = c_i + \frac{1 - c_i}{1 + \exp[-a_i(\theta - b_i)]}$$

where $P(Xi=1|\theta)$ is the probability of a correct response to item i, c_i is the guessing parameter, a_i is the discrimination parameter, b_i is the difficulty parameter, and θ is the person's latent trait level.

Latent Trait (θ) represents the underlying ability or trait of an individual that the test is intended to measure. In educational testing, it might represent a student's proficiency in a subject (Baker & Kim, 2017). The model estimates three parameters for each item on the test: (1) discrimination (a) which indicates how well an item differentiates between individuals with different levels of the latent trait. Higher values of the discrimination parameter suggest that the item is better at distinguishing between individuals with slightly different abilities. (2) Difficulty (b) which reflects the level of the latent trait required for a 50% chance of answering the item correctly. A higher difficulty parameter means that the item is more challenging. (3) Guessing (c) which accounts for the probability that an individual with a very low level of the latent trait would still answer the item correctly by guessing. It is particularly relevant for multiple-choice items where even those with little knowledge have a non-zero chance of guessing the correct answer (de Ayala, 2022; Hambleton et al., 1991).

The 3PL IRT model is important for several reasons, particularly in fields like educational testing, psychological assessments, and survey research. Its significance stems from its ability to provide a more nuanced and accurate representation of how test items function and how individuals' abilities are measured. The 3-PL IRT model is important because it offers a sophisticated and accurate approach to measuring latent traits, particularly in contexts where guessing is a factor. Its ability to account for guessing, combined with detailed item analysis and improved measurement precision, makes it a powerful tool for developing, analyzing, and administering tests in a fair and effective manner (de Ayala, 2022; Hambleton et al., 1991).

2.1 Advantages of the 3PL Model

2.1.1 Handling Guessing Behavior. One of the main advantages of the 3-PL model is its ability to account for guessing. In multiple-choice tests, especially, even individuals with low ability have a non-zero chance of answering an item correctly by guessing. The 3-PL model incorporates this by introducing a "guessing" parameter (ccc), which estimates the probability of a correct response purely by chance. This makes the model more realistic and accurate in scenarios where guessing is likely (Hambleton et al., 1991; Korompot et al., 2024).

2.1.2 More Detailed Item Information. By estimating three parameters—difficulty, discrimination, and guessing—the 3-PL model provides a richer analysis of test items. This

helps test developers and psychometricians understand not just how difficult an item is, but also how well it discriminates between individuals with different levels of ability and how likely individuals are to guess correctly. This detailed information can be used to improve the quality and fairness of tests.

2.1.3 Improved Measurement Precision. The model allows for a more precise measurement of an individual's latent trait (such as ability) across a range of ability levels. By accounting for guessing and varying discrimination levels, the 3-PL model can produce more accurate estimates of a person's ability, particularly for individuals at the extremes of the ability distribution (very high or very low ability).

2.1.4 Item Selection and Test Adaptation. In computer adaptive testing (CAT), where the test adapts in real time based on the test-taker's performance, the 3-PL model is particularly useful. It helps in selecting items that are appropriate for the test-taker's ability level, while also considering the likelihood of guessing and the discrimination power of items. This leads to more efficient testing, where fewer items are needed to achieve accurate measurements.

2.1.5 Reducing Bias. By considering the guessing behavior, the 3-PL model reduces potential biases in ability estimates that could occur if guessing were ignored. This contributes to the fairness and validity of the test, as it helps ensure that the test scores reflect true ability rather than random guessing.

2.1.6 Applicability Across Different Tests. The 3-PL model can be applied to a wide range of test types and formats, making it a versatile tool in psychometrics. Its flexibility in handling different item characteristics makes it valuable in diverse testing situations, from academic assessments to psychological evaluations (Hambleton & Swaminathan, 1985).

2.1.7 Benchmarking Against Simpler Models. The 3-PL model can be used as a benchmark to evaluate whether simpler models (like the 1-PL or 2-PL models) are sufficient for a particular testing scenario. If guessing is a significant factor, the 3-PL model will likely provide a better fit to the data, justifying its use over simpler models (Hambleton et al., 1991; Hambleton & Swaminathan, 1985).

The purpose of the present study is o examine the MC items of PIRLS 2016 using the 3PL model. Atmawinata et al. (in press), examined PIRLS items using the 2PL IRT model but ignored the guessing that MC items may generate. In this study we attempt to focus on the MC items with the 3PL model and evaluate to what extent guessing is at play.

3. Method

31. Data

Publicly available data from PIRLS 2016 for the USA was used for this analysis. Multiple-matrix sampling where a large number of items is distributed in several test booklets to maximize the tested content without overwhelming individual examinees is used in PIRLS assessments (see Baghaei & Robitzsch, 2025). The 175 items of PIRLS 2016 were distributed in 16 different test booklets each containing two tasks or passages with associated items. To establish a link between the booklets, each passage (and its relevant items) appeared in three different booklets. This strategy connected all the booklets with some shared items and enabled joint calibration of all items. The 21 MC items included in Form 1 were used in this analysis. A total of 272 American 4th graders had taken test Form 1.

3.2 Analysis and Results

The 3PL IRT model was applied to the 21 MC items of Form 1 of PIRLS 2016. The R package (R Core Team, 2022) *mirt* (Chalmers, 2012) was used to estimate the model. In the first step, the overall fit of the data to the 3PL model was evaluated using the *M2* statistic (Maydeu-Olivares & Joe, 2005). The *M2* statistic was nonsignificant which is a sign that there is not much difference between the model and the data (*M2*=179.81, *df*=168, *p*=.25, *RMSEA*= .01, *CFI*=.99, *TLI*=.98). This was also corroborated with high values of *CFI* and *TLI* (>.95) and a low value of *RMSEA* (<.08). In other words, the model, overall, fits the data.

Table 1 shows the parameters of the 3PL IRT model. '*a*' is the discrimination parameter and shows how well item discriminates between high-ability examinees and low-ability examinees. Larger discrimination indicates steeper item characteristic curve and a sign that the item has a stronger relationship with the latent trait. It can be compared with item loading in factor analysis. The discrimination values range from .46 to 2.62. The column r_{pb} shows the point-biserial correlation between the items and the rest scores, i.e., the total score minus the item under consideration (corrected item-total correlation) which, in the classical test theory, is an index of discrimination. Comparison between r_{pb} and the IRT a parameter shows that while they agree to a great extent, there are some inconsistencies as well. Item 6 has the lowest discrimination based on both r_{pb} and *a*-parameter. However, Item 11 which has the highest *a*-parameter has a very small r_{pb}. Items 13 and 20 with *a*-parameters of 1.11 and 1.17, respectively have r_{pb} values of .37 and .27, respectively. This means that a higher *a*-parameter does not always correspond with a higher item-total correlation.

Under the 1PL and 2PL models, the location or difficulty parameter '*b*' shows the ability level that is associated with 50% chance of getting the item right. Larger values indicate more difficult items. However, in 3PL model, since the lower asymptote of the ICC is not zero and is '*c*', the definition of item difficulty changes. In 3PL, item difficulty is the ability location where the probability of getting the item correct is halfway between the value of '*c*' and 1. The guessing parameter is the probability of answering an item correctly by guessing alone with no knowledge of the subject. For example, if the guessing parameter for an item is .30, it means the chances that an examinee answers the item correctly just by chance is 30%. The 'c' parameter ranges between 0 and 1; values smaller than .35 are considered acceptable (Baker & Kim, 2017). The difficulty parameters '*b*' here range from –1.89 to .50 and the guessing parameters '*c*' range from 0 to .72. Two items with labels 'R11F05M' and 'R41I01C'have unacceptable guessing parameters.

Table 1 also shows the item outfit and infit mean square values. Outfit and infit mean square values should be between .50 to 1.50 (Abdullaeva et al., 2024; Linacre, 2023). As Table 1 shows all the items have acceptable outfit and infit statistics.

Table 1
3PL IRT Item Parameters

Item	Label	*a*	*b*	*c*	Outfit	Infit	r_{pb}
1	R11F01M	1.73	-0.79	0.29	0.872	0.928	.41
2	R11F02M	0.53	-1.9	0	1.219	1.095	.21
3	R11F03M	1.85	-1.18	0.07	0.78	0.897	.45
4	R11F04M	1.62	-1.34	0.16	0.845	0.921	.40
5	R11F05M	2.33	-0.78	0.4	0.792	0.935	.39
6	R11F06C	0.46	-1.32	0.01	1.189	1.144	.17

7	R11F08C	1.65	-0.19	0	0.863	0.886	.50
8	R11F10C	1.03	-1.12	0.01	0.981	0.974	.37
9	R11F11M	0.64	0.25	0.21	1.171	1.117	.20
10	R11F13M	0.92	-0.66	0	1.017	1.021	.33
11	R41I01C	2.63	-0.46	0.72	0.907	1.032	.24
12	R41I02M	0.66	-0.41	0	1.133	1.082	.26
13	R41I05M	1.11	-1.03	0	1.006	0.965	.37
14	R41I06M	1.69	0.12	0.28	0.988	0.999	.35
15	R41I08M	2.16	-1.77	0	0.608	0.876	.42
16	R41I09C	1.07	-0.9	0	1.034	0.974	.38
17	R41I10M	0.86	0.01	0	1.023	1.019	.32
18	R41I12M	1.22	-0.58	0	0.918	0.938	.42
19	R41I13C	1.1	-0.44	0.06	1.003	0.988	.36
20	R41I14C	1.17	0.23	0.32	1.082	1.064	.27
21	R41I15C	0.84	0.5	0	1.045	1.040	.28

Table 2 shows the S_X2 fit statistics (Orlando & Thissen, 2000) for the items, their associated degrees of freedom, RMSEA, and *p*-values. Items with statistically significant S_X2 ($p<.05$) have a poor fit and are candidates for removal. Table 2 indicates that items 1 and 15 have poor fits.

Table 2
S_X2 Item Fit Values

Item	**Label**	**S_X2**	**df.S_X2**	**RMSEA.S_X2**	**p.S_X2**
1	R11F01M	20.86797	11	0.057	0.034
2	R11F02M	8.376538	12	0	0.755
3	R11F03M	6.562946	10	0	0.765
4	R11F04M	9.34403	11	0	0.590
5	R11F05M	7.139156	9	0	0.622
6	R11F06C	8.145569	13	0	0.833
7	R11F08C	2.910517	9	0	0.967
8	R11F10C	3.894154	12	0	0.985
9	R11F11M	14.16104	13	0.018	0.362
10	R11F13M	9.254462	12	0	0.681
11	R41I01C	4.900278	9	0	0.842
12	R41I02M	13.81279	13	0.015	0.387
13	R41I05M	16.64506	12	0.037	0.163
14	R41I06M	8.052997	12	0	0.780
15	R41I08M	13.6554	6	0.068	0.033
16	R41I09C	13.63379	12	0.022	0.324
17	R41I10M	16.41406	11	0.042	0.126
18	R41I12M	10.13756	11	0	0.518
19	R41I13C	20.29611	12	0.050	0.061
20	R41I14C	14.52141	12	0.027	0.268

21	R41I15C	4.88534	11	0	0.936

Figure 1 displays the item characteristic curves for the 21 PIRLS items. The lower ends of the ICCs indicate the guessing parameters and their slope shows their discrimination. For example, the lower end of item 5 is very high which shows that this item has a high discrimination parameter while items 7 and 8 have guessing parameters of zero. The slope of the ICCs indicates the item discrimination. Items 6 and 9 have relatively flat which is a sign of low discrimination while items 2 and 15 are highly discriminating. Item 11 has a relatively flat ICC but has the highest item discrimination parameter (a=2.62). This item has the highest guessing parameter as well (c=.72). This means that when an examinee has a 72% probability of getting an item correct only by chance, the item discrimination does not make much sense. The advantage of examining ICCs which show the behavior of the items graphically displays issues which are not easily visible by checking the numerical values.

Figure 1

Item Characteristic Curves

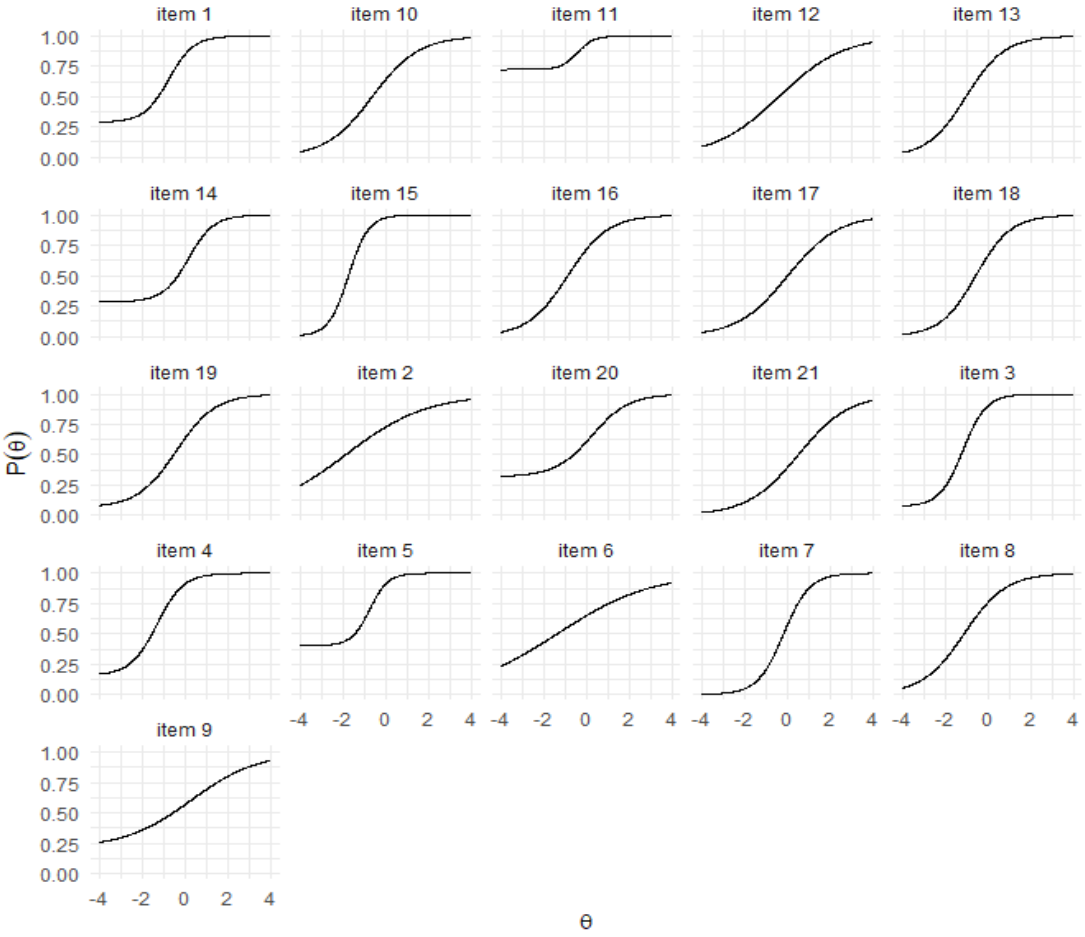
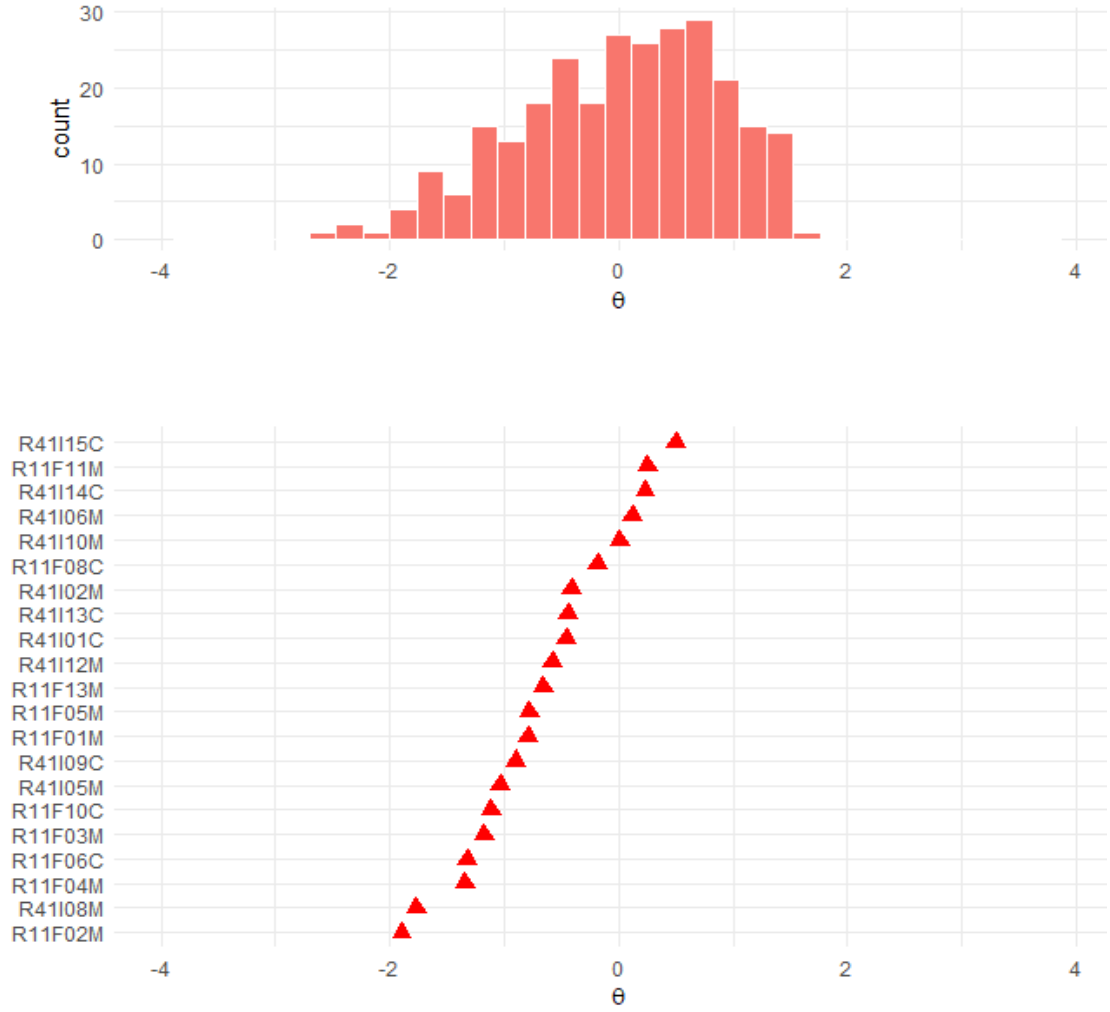


Figure 2 is the item-person map of the test. The figure in the upper part shows the distribution of the ability parameters and the figure in the lower part shows the locations of the items. The expectation is that the item locations (red arrows) cover the entire range of the ability distribution. As Figure 1 shows, the ability parameters range approximately from –2.60 to 1.70. However, as the lower figure displays, the item parameters range from approximately -2 to .50.

That, the items do not cover the higher range of the ability scale and this test needs some harder items.

Figure 2

Item-Person Map



4. Discussion

The analysis of the PIRLS 2016 data from 4th graders in the USA provides valuable insights into the psychometric properties of the test items using the 3PL IRT model. This discussion will explore the implications of the findings, address potential limitations, and compare them with findings from other studies examining item response theory models in large-scale assessments. The application of the 3PL IRT model in this study provided a deeper understanding of the multiple-choice (MC) items used in PIRLS 2016, particularly regarding guessing behavior, discrimination power, and item-person targeting. This study's findings align with previous research but also highlight some unique aspects of test design challenges in large-scale educational assessments.

A relevant study by Atmawinata et al. (in press) analyzed PIRLS 2021 data from Kazakhstan using a 2PL IRT model, which does not account for the guessing parameter. Their findings showed that most items had acceptable discrimination values but noted potential item misfit issues due to item difficulty inconsistencies. Our study expands on their findings by demonstrating that ignoring the guessing parameter may lead to overestimating item quality, particularly in multiple-choice formats, where some items may be answered correctly by

chance rather than ability. This underscores the importance of using the 3PL model, especially in assessments where guessing is a significant factor.

Similarly, Korompot et al. (2024) conducted an IRT-based gender DIF analysis in the B2 First Exam's reading comprehension section. Their study found that some items exhibited higher guessing parameters for specific subgroups, potentially skewing the assessment results. This resonates with our study's observation that two items in PIRLS 2016 had unacceptably high guessing parameters, which could impact score validity. While their study focused on group differences, our findings emphasize the broader issue of test reliability and construct validity when guessing is present.

In contrast, Baghaei and Effatpanah (2024) applied nonparametric IRT to Likert-type assessments and found that difficulty and discrimination were well-modeled, but traditional IRT models often underestimated random guessing effects. This further validates our finding that item characteristic curves (ICCs) offer a clearer depiction of item behavior, particularly in pinpointing items where guessing inflates performance metrics. While their work focuses on Likert-scale data, the overarching concern regarding guessing distortion applies across multiple test types.

The findings of this study have several important implications for test design, educational assessment policy, and psychometric research. The identification of high-guessing items suggests that multiple-choice items should be carefully reviewed to ensure they measure actual reading ability rather than random selection of answers. One approach could be the incorporation of more constructed-response items or the use of polytomous scoring models, which reduce the influence of guessing. The item-person map revealed that PIRLS does not effectively assess students at the highest ability levels, which is consistent with Linacre (2023), who emphasized the importance of ensuring item difficulty covers the full ability range. This suggests that future versions of PIRLS should include more challenging items to ensure higher-achieving students are adequately evaluated. Hambleton et al. (1991) argued that multiple-choice formats are efficient but prone to guessing, whereas constructed-response items provide richer insights into student ability. Our findings reinforce this by showing that some PIRLS MC items may not fully capture students' reading comprehension skills. A hybrid model that balances MC and open-ended questions could enhance measurement accuracy. Baker & Kim (2017) emphasized that ICCs are crucial tools for assessing item behavior. This study confirms that graphical representations can reveal misfitting items that numerical statistics alone might overlook. This highlights the need for test developers to incorporate ICC analyses in routine test evaluations.

Since PIRLS is an international assessment, it would be beneficial to apply the 3PL IRT model to other participating countries to examine whether similar guessing issues persist across different education systems. This could help determine whether test design flaws are systemic or country-specific. Future research could also explore whether Bayesian IRT models (e.g., Effatpanah & Baghaei, 2023) provide better item parameter estimates than classical 3PL approaches. Bayesian methods may reduce parameter uncertainty, particularly in small-sample scenarios.

As computer-based testing (CBT) becomes more prevalent, it would be interesting to examine whether TEIs (e.g., interactive reading tasks, adaptive questioning) could help mitigate guessing effects and provide a more authentic reading assessment. Research on computer-adaptive testing (CAT) in PIRLS would also be valuable. Given findings from Korompot et al. (2024) on gender-related DIF, further analysis could explore whether certain

PIRLS MC items exhibit bias towards specific student subgroups (e.g., gender, socioeconomic status, or language background). Since PIRLS also collects teacher and school questionnaire data, future studies could link item-level performance to instructional practices, helping identify whether certain teaching strategies lead to better test performance.

5. Conclusion

The psychometric analysis of PIRLS 2016 multiple-choice items using the 3PL IRT model contributes to a better understanding of the test's validity and reliability. While the test demonstrates strong overall model fit, the findings suggest that guessing remains a concern and that item targeting could be improved. Comparisons with similar studies confirm that these challenges are not unique to PIRLS but are common across large-scale educational assessments. Addressing these issues—through better item design, diverse question formats, and new psychometric methodologies—will be critical in enhancing future iterations of PIRLS and similar reading assessments.

Acknowledgments
We sincerely thank the anonymous reviewers for their valuable comments and suggestions, which have greatly improved the quality of this paper.

Declaration of Conflicting Interests
The authors declare that there are no conflicts of interest related to this work.

Declaration of Applying AI
AI was not used for the preparation of this manuscript.

Funding
The authors did not receive any specific grant or funding for this research.

References

Abdullaeva, B. S., Abdullaev, D., Khursanov, N. I., Kadirova, K. B., & Djuraeva, L. (2024). Modelling local item dependence in cloze tests with the Rasch model: Applying a new strategy. *International Journal of Language Testing, 14*, 97–103. DOI: 10.22034/ijlt.2024.435779.1319

Atmawinata, M. R., Herwina, W., Bulan, S., Wikanengsih, W., Darheni, N., & Tashtemirova, G. (Advance online publication). Item response theory analysis of the Progress in International Reading Literacy Study (PIRLS) 2021 in Kazakhstan. *International Journal of Language Testing*. DOI: 10.22034/ijlt.2024.456241.1343

Baghaei, P., & Effatpanah, F. (2024). Nonparametric kernel smoothing item response theory analysis of Likertitems. *Psych*, *6*(1), 236–260. https://doi.org/10.3390/psych6010015

Baghaei, P., & Robitzsch, A. (2025). A tutorial on item response modeling with multiple groups using TAM. *Educational Methods & Psychometrics, 3*, 14. https://dx.doi.org/10.61186/emp.2025.1

Baker, F., & Kim, S-H. (2017). *The basics of item response theory using R*. Springer Cham.

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

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.

de Ayala, R. J. (2022). *The theory and practice of item response theory (2nd Ed.)*. Guilford Press.

Effatpanah, F., & Baghaei, P. (2022). Exploring rater quality in rater-mediated assessment using the nonparametricitem characteristic curve estimation. *Psychological Test and Assessment Modeling, 64*(3), 216–252. https://www.psychologie-aktuell.com/fileadmin/Redaktion/Journale/ptam_2022-3/PTAM__3-2022_2_kor.pdf

Effatpanah, F., & Baghaei, P. (2023). Kernel Smoothing Item Response Theory in R: A Didactic. *Practical Assessment, Research & Evaluation*, *28*, 7. https://doi.org/10.7275/pare.1261

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage.

Hambleton, & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Springer Dordrecht.

Korompot, C. A., Siregar, I., Khursanov, N. I., Abdullaev, D., & Mohamed, K. M. (2024). Investigating gender DIF in the reading comprehension section of the B2 First Exam. *International Journal of Language Testing, 14*, 57–66. DOI: 10.22034/ijlt.2023.421011.1301

Linacre, J. M. (2023). *Winsteps® Rasch measurement computer program User's Guide*. Version 5.6.0. Portland, Oregon: Winsteps.com

Maydeu-Olivares A., Joe H. (2005). Limited- and full-information estimation and goodness-of-fit testing in 2^n contingency tables: A unified framework. *Journal of the American Statistical Association*, *100*, 1009–1020. https://doi.org/10.1198/016214504000002069

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2017). *PIRLS 2016 international results in reading*. Retrieved from Boston College, TIMSS & PIRLS International Study Center website: http://timssandpirls.bc.edu/pirls2016/international-results/

Mullis, I. V. S., & Martin, M. O. (Eds.). (2015). *PIRLS 2016 assessment framework* (2nd Ed.). Retrieved from Boston College, TIMSS & PIRLS International Study Center website: https://timssandpirls.bc.edu/pirls2016/framework.html

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24*(1), 50–64. https://doi.org/10.1177/01466216000241003

R Core Team (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.