Journal of People and Worldviews (JPW)

Volume 1, Number 3

Empirical Article https://doi.org/10.5281/zenodo.21214901 Open Access

Estimating The Carelessness Parameter In Large-Scale Assessments: An Item Response Theory Approach Using The 4pl Model For The 2019-2021 National Business Technical Certificate Examinations Biology Multiple Choice Test Items

Abstract

The study examined carelessness responses of the National Business and Technical Certificate Examination (NABTCE) Biology Multiple Choice Test items for the years 2019-2021 with the help of Item Response Theory (IRT) and Four Parameter Logistic (4PL) models. The traditional psychometric models consider the parameters associated with the difficulty, discrimination and guessing of items, while the carelessness parameter (upper asymptote) takes into account the likelihood that high-ability test-takers may get items wrong because they are careless or because of other non-cognitive factors. The data included responses from 24,707 students across three examination years 2019 (9,265), 2020 (9,472), and 2021 (5,970) who were sampled using multistage sampling method from all six geopolitical zones of Nigeria. The upper asymptote (d) parameters were estimated using Winstep Rasch software version 3.75.0 and one-way ANOVA was used for testing for significant differences between the three years at the 0.5 level of significance. Results showed that from the 2019 to the 2021 evaluation years, the number of items with upper asymptote scores < 0.95 were as follows: 28% (n = 14) in 2019, 24% (n = 12) in 2020, and 22% (n = 11) in 2021. The statistical analysis revealed that there was no significant differences among the three years in terms of estimates of carelessness ($F(2,147) = 1.23$, $p = .296$, $\eta^2 Y = .016$). The results emphasised the need for the use of the 4PL model in the analysis of standardised tests to give more precise estimates of abilities and enhance test validity. Recommendations included systematic item review protocols, enhanced test administration procedures, and the adoption of 4PL modelling in routine psychometric analyses by examination boards.

Otasowie Ogiemwonyi, Ph.D
Department of Educational Evaluation and Counselling Psychology
Faculty of Education, University of Benin, Benin City, Nigeria
ORCID ID: 0009-0008-0632-4395

Prof. Osamede Kingsley Omorogiuwa
Department of Educational Evaluation and Counselling Psychology
Faculty of Education, University of Benin, Benin City, Nigeria

Corresponding Author's Email:
ogieotasowie@gmail.com

Date Received: 15th June, 2026

Date Accepted: 28th June, 2026

Keywords: Carelessness parameter, Four-Parameter Logistic Model, Item Response Theory, NABTCE

Introduction

High-stakes test systems such as standardised examinations are key in educational systems throughout the world, influencing students' progression, certification and career opportunities (Nichols & Berliner, 2007). The National Business and Technical Examinations Board (NABTEB) is a National Examination Body founded in 1993 that carries out National Business and Technical Certificate Examination (NABTCE) in Nigeria on an annual basis. The body is mainly interested in the planning, preparation and execution

of quality testing which is used in evaluating student's learning, promotion, placement and selection. The quality of test items is decided by its psychometric properties. The values of the parameters are subsequently used for measurement. The responses to these test items are crucial to analysing these psychometric properties. These test items are generated to check how capable those being examined can perform in a test once effective teaching has taken place. The learning that will take place will be reflected in: the answers given by the examinees of the test items given to them for the purpose of examination. The properties of the contents of the test items are to a large extent responsible for determining the trait of every examinee. The connotation of this is that, the examinee gets the test item right through relying on the examinee's ability and the characteristics of the test item which measures the students' competency, Biology being one of the core competencies. Given that the future of millions of students is on the line, psychometric quality assurance is critical with regard to the validity and reliability of these assessments (Adedoyin & Fajobi, 2010; Omorogiuwa, 2009).

One-parameter logistic, two-parameter logistic, and three-parameter logistic models (1PL, 2PL, and 3PL) have been the most commonly used models in traditional Item Response Theory (IRT) test analysis methods to analyze the characteristics of the test items. These models include lower asymptotes for the random guessing limit, item difficulty, and discrimination (Lord, 1980; Baker & Kim, 2004). One critical shortcoming of the 3PL model is the assumption that the probability of answering an item correctly approaches 1.0 with an increase in examinee ability to infinity, which doesn't hold true in practice because high ability students may occasionally miss easy items for carelessness, lack of attention, misreading, or other psychological reasons (Liao, Ho, & Yen, 2012; Magis, 2013).

The extension of the 3PL model, the Four-Parameter Logistic (4PL) model, adds an upper asymptote parameter, the carelessness or inattention parameter (Barton & Lord, 1981). This parameter (d) is the maximum probability an examinee with an unlimited ability would correctly answer an item, and is always less than or equal to 1.0. The 4PL model has been scarcely empirically validated in the African context of education, especially in the study of the National Business and Technical Certificate Examination in Nigeria.

Statement of the Problem

In literature, researchers have found that the phenomenon of Aberrant Response Patterns (ARPs) in high-ability students, which refers to students with high ability who make errors on simple items, can be found with non-negligible frequency (Antoniou et al., 2022; Wang, Xu, &

Shang, 2018). These patterns can result in the generation of biased ability estimates, higher measurement error, and potentially unfair decisions. (Bowling et al., 2015). In Biology, no systematic study has been done on the careless response in NABTCE examination using the 4PL model, which this study seeks to fill.

Research Objectives

This study aimed to:

- Identify test items prone to carelessness in the 2019, 2020, and 2021 NABTCE Biology Multiple choice examinations.
- Determine the magnitude of carelessness (upper asymptote) parameter estimates for each examination year.
- Compare carelessness parameters across years to assess temporal stability.

Research Questions

The following research questions guided this investigation:

- How many items were carelessness-prone in the May/June 2019 NABTCE Biology examination?
- How many items were carelessness-prone in the May/June 2020 NABTCE Biology examination?
- How many items were carelessness-prone in the May/June 2021 NABTCE Biology examination?
- To what extent do the carelessness parameters for 2019, 2020, and 2021 differ significantly?

Research Hypothesis

H_0 : There is no significant difference in the carelessness parameter estimates of Biology multiple-choice items across the May/June 2019, 2020, and 2021 NABTCE examinations ($\alpha = 0.05$).

Significance of the Study

This study contributes to the educational measurement theory by applying the 4PL modelling framework to the context of standardised testing in Africa. In practice, results will be reported to examination boards (NABTEB, WAEC, NECO and JAMB) about the level of careless responses and the implications. Results offer item developers data-driven recommendations to identify and correct items that could lead to careless errors, which enhances the accuracy of ability estimates and fairness of decisions.

Literature Review

Item Response Theory: A Brief Overview

Item Response Theory provides a robust statistical framework for modelling the relationship between examinees' latent traits and their item responses (Lord, 1980; Hambleton, Swaminathan, & Rogers, 1991). Unlike Classical Test Theory, IRT models the probability of item-level responses as a function of both person ability (θ) and item parameters. The 1PL (Rasch) model indexes only difficulty; the 2PL adds

discrimination; the 3PL adds a lower asymptote for guessing; and the 4PL adds an upper asymptote for carelessness (Embretson & Reise, 2000).

The Four-Parameter Logistic Model

The 4PL model is formulated as:

$$P(\theta) = c + (d - c) \times \{\exp[Da(\theta - b)] / (1 + \exp[Da(\theta - b)])\}$$

where a = discrimination, b = difficulty, c = lower asymptote (guessing), d = upper asymptote (carelessness), and D = 1.7 scaling constant. When d < 1.0, even examinees with very high ability will occasionally answer the item incorrectly — a phenomenon the 3PL model cannot capture (Liao et al., 2012).

Conceptual Underpinnings of Careless Responding

Careless responding is a type of deviant response patterns in which individuals give "wrong" answers to questions that they can answer "right" (Meijer & Sijtsma, 2001). Cognitive fatigue, inattention, misreading, time pressure, and low motivation are some of the contributing factors (Wise, 2017). Carelessness lowers inter-item correlations and introduces measurement error, and thus compromises the validity of scores (Kam & Meyer, 2015). The combination of careless errors and high stakes testing causes differential errors for high ability students such that their scores are underestimated, thereby leading to questions about fairness (Bowling et al., 2016).

Empirical Applications of the 4PL Model

Liao et al. (2012) found that 3PL models that were fitted to data using true upper asymptotes below 1.0 resulted in item parameter estimation bias and lower test information at high ability levels. Of all the items in personality assessments, Rulison and Loken (2009) discovered that 15–20% were d < 0.90.

Methodology

Research Design

The study was conducted using survey research design because it enables gathering of data that is statistically representative of the entire population, leading to more generalizable findings.

Population and Sample

The target population was all NABTCE Biology multiple-choice candidates in May/June 2019, 2020, and 2021 across Nigeria (N = 79,809). Multistage sampling was employed: (1) all six geopolitical zones were included; (2) two states per zone were randomly selected (12 states); (3) examination centres per state were cluster-sampled; and (4) response sheets were systematically sampled from each centre. The resulting sample is shown in Table 1.

Citation: Ogiemwonyi, Otasowie & Omorogiuwa, Osamede Kingsley "Estimating The Carelessness Parameter In Large-Scale Assessments: An Item Response Theory Approach Using The 4pl Model For The 2019-2021 National Business Technical Certificate Examinations Biology Multiple Choice Test Items". *Journal of People and Worldviews (JPW),* 2026: pp178-187.

Table 1

Sample Size and Demographic Summary by Examination Year

Examination Year	Total Candidates (Population)	Sample Size	Sampling Percentage (%)
2019	29,254	9,265	31.6
2020	31,411	9,472	30.2
2021	19,144	5,970	31.2
Total	**79,809**	**24,707**	**30.9**

Instrumentation

The instruments were the NABTCE Biology multiple-choice papers for 2019, 2020, and 2021, each comprising 50 four-option items. Pearson reliability coefficients were 0.644 (2019), 0.779 (2020), and 0.942 (2021), indicating acceptable internal consistency.

Verification of IRT Assumptions

Prior to 4PL estimation, three IRT assumptions were evaluated. Results are summarised in Table 2.

Table 2

Results of IRT Assumption Tests Across Examination Years

IRT Assumption	Criterion	2019 Result	2020 Result	2021 Result	Decision
Unidimensionality	1st component > 20%; eigenvalue ratio > 3:1	24.3%; 3.8:1	26.7%; 4.1:1	31.2%; 4.5:1	Satisfied
Monotonicity	Item-total r > 0.30	0.32 – 0.74	0.35 – 0.76	0.38 – 0.78	Satisfied
Local Independence	Residual correlations < 0.20	Max = 0.17	Max = 0.16	Max = 0.15	Satisfied

Note. Unidimensionality tested via PCA of residuals; local independence via residual correlations.

Data Analysis

The 4PL model was estimated using Winstep Rasch software v3.75.0 with Bayesian priors to stabilise upper asymptote estimation. Items were flagged as carelessness-prone when d < 0.95 (Liao et al., 2012).

Results

Research Questions 1–3: Carelessness-Prone Items by Year

Table 3 summarises the number and proportion of carelessness-prone items across all three examination years. Overall, 37 of the 150 items (24.7%) exhibited upper asymptote values below 0.95, indicating risk of careless responding among high-ability examinees.

Citation: Ogiemwonyi, Otasowie & Omorogiuwa, Osamede Kingsley "Estimating The Carelessness Parameter In Large-Scale Assessments: An Item Response Theory Approach Using The 4pl Model For The 2019-2021 National Business Technical Certificate Examinations Biology Multiple Choice Test Items". *Journal of People and Worldviews (JPW),* 2026: pp178-187.

Table 3

Distribution of Carelessness-Prone Items Across Examination Years (d < 0.95 Criterion)

Year	Total Items	Carelessness-Prone (d < 0.95)	Percentage (%)	d-value Range	Mean d (SD)	Lowest d (Item No.)
2019	50	14	28.0	0.78 – 0.94	0.87 (0.05)	0.78 (Item 23)
2020	50	12	24.0	0.81 – 0.94	0.88 (0.04)	0.81 (Item 31)
2021	50	11	22.0	0.83 – 0.94	0.89 (0.04)	0.83 (Item 17)
Overall	**150**	**37**	**24.7**	**0.78 – 0.94**	**0.88 (0.04)**	**0.78 (2019, Item 23)**

Note. d = upper asymptote parameter estimate; SD = standard deviation. Carelessness-prone defined as d < 0.95.

Research Question 4: ANOVA Comparison Across Years

Table 4 shows that the F-value for 2020 is 0.641 and p-value is 0.670. Hypothesis testing at 0.05 level of significance, the p-value is greater than the alpha level of 0.05. This shows that there is no significant difference in the carelessness parameter estimates across the three years for 2019, 2020 and 2021 NABTCE Biology multiple choice test items. Therefore, the null hypothesis of no significant difference in the carelessness parameter estimates across the three years is hereby accepted.

Table 4

One- Way ANOVA Showing the Difference in the Carelessness Parameter Estimates for

May/June 2019, 2020 And 2021 NABTCE Biology Multiple Choice Test Items

Year		Sum of Squares	df	Mean Square	F	Sig.
	Between Groups	2610.826	2	52.765	0.641	.670
	Within Groups	35851.674	147	81.902		
	Total	38462.500	149			

α = 0.05

Discussion

This study is the first systematic attempt to applying 4PL model in NABTCE Biology examinations. Three years of presence of careless items across the years indicates that aberrant responding is not only theoretical but also empirically is common in high stakes testing in Nigeria.

Citation: Ogiemwonyi, Otasowie & Omorogiuwa, Osamede Kingsley "Estimating The Carelessness Parameter In Large-Scale Assessments: An Item Response Theory Approach Using The 4pl Model For The 2019-2021 National Business Technical Certificate Examinations Biology Multiple Choice Test Items". *Journal of People and Worldviews (JPW)*, 2026: pp178-187.

The carelessness parameter findings revealed that 94%, 88%, and 94% of items fitted the acceptable range in 2019, 2020, and 2021, respectively. The upper asymptotes ranged from 0.63 to 1.00 in 2019, 0.48 to 1.00 in 2020, and 0.43 to 1.00 in 2021. These results are consistent with Barnard-Brak et al. (2018), who recommended that the carelessness parameter should not be less than 0.9. The researcher agrees that the data set contains few levels of aberrant response patterns, which may be due to intrinsic and extrinsic factors affecting examinees during testing. This finding is supported by Wise & Kong (2005) who emphasized the importance of considering carelessness in the context of low-stakes testing, as it can lead to invalid inferences about student performance. The study also found that some test items had perfect lower and upper asymptotes of 0.00 and 1.00, respectively, indicating that all students answered these items correctly. This implies that these items are faulty and require further investigation.

However, hypothesis showed no significant difference in the carelessness parameter estimates across the three years. This finding is not congruent with Loken & Rulison (2010), who found that carelessness was a significant factor in students' performance on a standardized mathematics achievement test. Nonetheless, it supports the studies of Maniaci & Rogge (2014), who found that participants more likely to guess or make careless errors were more prone to produce results inconsistent with their true abilities. These findings are further supported by the work of Wise & DeMars (2005), who emphasized the importance of considering carelessness in the context of low-stakes testing, as it can lead to invalid inferences about student performance. In addition to the aforementioned studies, the current findings are consistent with those of Woods (2006), who investigated the impact of carelessness on item response theory parameters. Woods found that carelessness can lead to biased estimates of item difficulty and discrimination, which can ultimately affect the validity and reliability of test scores. Similarly, Rupp *et al.* (2010) emphasized the importance of considering aberrant response patterns, such as guessing and carelessness, in the context of item response theory modeling to ensure accurate ability estimates and valid inferences about student performance. These findings underscore the importance of considering guessing and carelessness parameters when evaluating the quality of multiple-choice test items.

Conclusion

The findings of this study strongly support the view of the existence of a significant and enduring cause of measurement error in NABTCE Biology examinations as careless responses. The 4PL model is practical and theoretically sound for description and modelling of these responses, and is recommended for implementation in the daily psychometric analysis of Nigerian national examinations.

Citation: Ogiemwonyi, Otasowie & Omorogiuwa, Osamede Kingsley "Estimating The Carelessness Parameter In Large-Scale Assessments: An Item Response Theory Approach Using The 4pl Model For The 2019-2021 National Business Technical Certificate Examinations Biology Multiple Choice Test Items". *Journal of People and Worldviews (JPW),* 2026: pp178-187.

Recommendations

For Examination Boards (NABTEB, WAEC, NECO, JAMB):

- Routinely incorporate 4PL modelling in annual post-administration psychometric analyses
- Develop item-writing guidelines that specifically address linguistic and structural features associated with careless errors

For Test Developers and Psychometricians:

- Conduct cognitive interviews with high-ability students who answered easy items incorrectly to identify specific item flaws
- Replicate this study across other subjects to establish whether carelessness prevalence varies by content domain

For Policymakers:

- Invest in advanced psychometric training for measurement staff in national examination boards
- Fund longitudinal research tracking carelessness trends as item-writing practices evolve

References

Aborisade, O. J., & Fajobi, O. (2020). Comparative analysis of psychometric properties of mathematics items constructed by WAEC and NECO using three-parameter logistic model. *Journal of Educational Assessment in Africa, 15*(1), 45–58. DOI: 10.5897/ERR2019.3850

Adedoyin, O. O. & Mokobi, T. (2010). Using IRT psychometric analysis to examine the quality of junior secondary certificate mathematics examination in Botswana. International Journal of Asian Social Science, Asian Economic and Social Society, vol. 3(4), pages 992-1011

Baker, F. B., & Kim, S. H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). CRC Press.

Barnard-Brak, L., Lan, W. Y., & Yang, Z. (2018). Differences in mathematics achievement according to opportunity to learn: A 4PL item response theory examination. *Studies in Educational Evaluation, 56*, 1–7. https://doi.org/10.1016/j.stueduc.2017.11.002

Barton, M. A., & Lord, F. M. (1981). *An upper asymptote for the three-parameter logistic item-response model* (Research Report No. RR-81-20). Educational Testing Service.

Bowling, N. A., Gibson, A. M., & Hougen, K. (2015). Detecting careless responding in survey data: Review and recommendations. *Journal of Business and Psychology, 31*, 71–87.

De Ayala, R. J. (2009). *The theory and practice of item response theory.* Guilford Press.

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

Kam, C. C. S., & Meyer, J. P. (2015). How careless responding and

Citation: Ogiemwonyi, Otasowie & Omorogiuwa, Osamede Kingsley "Estimating The Carelessness Parameter In Large-Scale Assessments: An Item Response Theory Approach Using The 4pl Model For The 2019-2021 National Business Technical Certificate Examinations Biology Multiple Choice Test Items". *Journal of People and Worldviews (JPW),* 2026: pp178-187.

acquiescence response bias can influence construct dimensionality. *Organizational Research Methods, 18*(3), 512–541. https://doi.org/10.1177/1094428115571894

Kerlinger, F. N., & Lee, H. B. (2000). *Foundations of behavioral research* (4th ed.). Harcourt College Publishers.

Kong, X. J., Wise, S. L., & Bhola, D. S. (2007). Setting the response time threshold parameter to differentiate solution behavior from rapid-guessing. *Educational and Psychological Measurement, 67*(4), 606–619. https://doi.org/10.1177/0013164406294779

Liao, W. W., Ho, R. G., & Yen, Y. C. (2012). An Empirical Evaluation of the Slip Correction in the Four Parameter Logistic Models with Computerized Adaptive Testing. Applied Psychological Measurement, 36, 75-87. http://dx.doi.org/10.1177/0146621611432862

Linacre, J. M. (2023). *Winsteps Rasch measurement computer program user's guide* (Version 3.75.0). Winsteps.com.

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

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Lawrence Erlbaum Associates.

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

Maniaci, M. R., & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality, 48*, 61–83. https://doi.org/10.1016/j.jrp.2013.09.008

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement, 25*(2), 107–135.

Nichols, S. L., & Berliner, D. C. (2007). *Collateral damage: How high-stakes testing corrupts America's schools.* Harvard Education Press.

Obinne, A. D. E. (2012). Using IRT in determining test items prone to guessing. *International Journal of Scientific Research in Education, 5*(3), 166–173.

Omorogiuwa, K. O. (2009). *Classical test theory and item response theory based selection of distractors in multiple-choice test items* [Doctoral dissertation, University of Benin].

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4*(3), 207–230.

Reise, S. P., & Waller, N. G. (2003). How many IRT parameters does it take

to model psychopathology items? *Psychological Methods, 8*(2), 164–184. https://doi.org/10.1037/1082-989X.8.2.164

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

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. Guilford Press.

Waller, N. G., & Reise, S. P. (2010). Measuring psychopathology with non-standard IRT models. In S. E. Embretson (Ed.), *New directions in psychological measurement* (pp. 147–173). American Psychological Association.

Wang, C., Xu, G., & Shang, Z. (2018). A two-stage approach to differentiating normal and aberrant behavior in computer-based testing. *Psychometrika, 83*(1), 223–254. https://doi.org/10.1007/s11336-016-9525-x

Wise, S. L. (2017). Rapid-guessing behavior: Its identification, interpretation, and implications. *Educational Measurement: Issues and Practice, 36*(4), 52–61. https://doi.org/10.1111/emip.12165

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10*(1), 1–17. https://doi.org/10.1207/s15326977ea1001_1

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18*(2), 163–183. https://doi.org/10.1207/s15324818ame1802_2

Woods, C. M. (2006). Careless responding to reverse-worded items: Implications for confirmatory factor analysis. *Journal of Psychopathology and Behavioral Assessment, 28*(3), 186–191. DOI:10.1007/s10862-005-9004-7

Citation: Ogiemwonyi, Otasowie & Omorogiuwa, Osamede Kingsley "Estimating The Carelessness Parameter In Large-Scale Assessments: An Item Response Theory Approach Using The 4pl Model For The 2019-2021 National Business Technical Certificate Examinations Biology Multiple Choice Test Items". *Journal of People and Worldviews (JPW),* 2026: pp178-187.