Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Enhancing Biochemistry Assessment Quality in Medical Education Through Item Response Theory (IRT)

Baharuddin Baharuddin (✉ **baharuddin@staff.ubaya.ac.id**)
Faculty of Medicine, Universitas Surabaya, Surabaya, Indonesia https://orcid.org/0000-0002-7079-5748

Lilis Handayani
Graduate School of Natural Science and Technology, Kanazawa University, Kanazawa, Japan
https://orcid.org/0000-0001-7210-1677

Rusli Rusli
SUPM Negeri Sorong/Politeknik Kelautan dan Perikanan Sorong, Sorong, Indonesia
https://orcid.org/0000-0001-5149-6438

Research Article

Keywords: solution, item response theory, assessment, biochemistry, medical education

Posted Date: January 16th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-3859753/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

Background

In medical education, particularly in biochemistry, crafting high-quality assessment questions is a primary challenge. Each item necessitates thorough evaluation, and precise identification of student abilities is crucial for maximally reflecting learning achievement.

Objective

This study aims to enhance assessment quality in biochemistry medical education by implementing Item Response Theory (IRT). This approach addresses Classical Test Theory (CTT) limitations. Recognizing the critical role of question quality in the learning process, the study investigates how IRT can more holistically and equitably assess student abilities. It includes a comparative analysis of student scores before and after IRT implementation.

Methods

Employing a mixed-method research approach, this study combines comparative quantitative analysis with qualitative ICC curve analysis in a pre-post experimental design. It focuses on biochemistry exam data from medical students (n = 89). IRT is used to measure the probability of student responses to questions, using parameters such as discrimination, difficulty level, and guessing probability. Jamovi software supports this analysis by accelerating computational processes.

Results

Significant improvements were observed in both question quality and student scores. Prior to IRT implementation, the average initial exam score was 56.1, which increased to 74.1 in the subsequent exam. The IRT evaluation indicated that the exam questions achieved a more effective differentiation between students of varying abilities. This improvement was evident from the increased person reliability and through Wright Map visualizations, which helped identify highly difficult questions via the Item Characteristic Curve (ICC).

Conclusion

The study advocates for integrating IRT as a standard method in biochemistry medical assessments. It highlights the necessity of assessments that are sensitive to individual student capabilities, providing more precise feedback for enhancing the quality of learning. These findings are crucial for evolving evaluation methodologies and advancing medical education standards.

Introduction

In medical education, particularly in biochemistry, creating high-quality assessment questions represents a significant challenge. Each item requires detailed evaluation through item analysis. Many students encounter difficulties in biochemistry learning, especially in the preclinical phase [1]–[3]. Attempts to improve this using Classical Test Theory (CTT) have been limited, as the analysis focuses only on the items without considering student ability and the probability of guessing. In short, CTT does not account for item-weighted assessment [4]. Therefore, a more comprehensive method is required, one of which is Item Response Theory (IRT). This method refines the CTT approach. However, research on the application of IRT in medical biochemistry is scarce, yet its findings could provide solutions for medical education units, especially those in charge of biochemistry, to improve question quality.

The uniqueness of the IRT model approach lies in its sensitivity to the individual's (subject's) response to each item [5], even including calculations for the probability of guessing for a population on an item. This approach utilizes the logit function and logistic parameters, such as discrimination (α), difficulty level (β), and pseudo-guessing parameter (c), to depict the relationship between items and various levels of a latent trait [6]. This method is particularly vital in examination systems like medical biochemistry, especially to enhance the quality of evaluation.

To measure the completeness of learning in medical biochemistry, an assessment-based evaluation is necessary. However, challenges often arise in evaluating exam results. Evaluations are sometimes solely based on student performance scores. Often, there is no item analysis evaluation, leading to situations where students are blamed for low scores, when in fact, the exam questions might be flawed in diction or too difficult compared to the taught material.

The application process of this method is straightforward, especially when using a single parameter (dichotomous), known as the Rasch Model. With technological advancements, this can be quickly executed using software like Jamovi. Implementing this method will assist in improving distractors in Multiple Choice Questions (MCQ) [7].

The research questions formulated in this study are:

- What is the range of theta performance values for assessment and students?
- Is there a significant change in the average score after implementing the revised questions?
- Does the theta value of items change when distractor modifications are made?

Methods

This study utilizes exam data to evaluate question performance [8]. The analytical approach consists of pre and post assessments following the implementation of Item Response Theory (IRT) to evaluate student performance scores. The evaluation was conducted on student assessments (n = 89) using two approaches. The first approach involved an analysis of average scores, and the second, a comparative quantitative analysis of item performance. The latter also included a qualitative study of the Item Characteristic Curve (ICC).

Items subjected to IRT were tabulated and visualized, presenting theta (θ) scores and their graphical changes in the ICC. To expedite the computational process in this research, Jamovi software with an open-access license was used. To eliminate subjectivity, this study was conducted after grades were published and no further grade disputes from students were present.

Results and Discussion

The evaluation was conducted using two approaches. The initial results showed a significant difference in average scores before and after the implementation of Item Response Theory (IRT). The items evaluated and improved using the IRT approach were visualized as follows. In this study, the overall reliability of the measurement tool was assessed, with a person reliability value of 0.780 for the first exam and 0.724 for the second exam. The analysis using IRT on exam data indicated a Person Reliability of 0.724 on a predetermined scale, suggesting that the exam had a good level of reliability in measuring the latent abilities or attributes of the participants.

There was a significant increase in scores before and after the implementation, with a $p\text{-value} < 0.001$ at an alpha of 0.05. The average score for the first exam was 56.1, and for the second exam, it was 74.1. Based on the score distribution, it was also observed that the exam questions effectively differentiated students (Fig. 1). Student abilities (scores) were well-identified across the low-middle-high spectrum, with the central tendency located in the middle area of the histogram curve distribution. These findings align with the study by Dorner et al. [9].

We identified items with very high difficulty levels. This identification was easily accomplished using the Wright Map distribution of the items. This visualization is adequately representative of item difficulty, as it computationally matches the items with the respondents' latent traits (ability to answer). It was evident that there are 11 particularly difficult items, numbered 20, 8, 46, 5, 39, 50, 7, 9, 29, 34, and 44. These items will subsequently be analyzed and improved upon after examining their Item Characteristic Curves (ICC).

ICC analysis is essential for examining three aspects. First, to ascertain the difficulty level of the items. Second, to assess the performance of the items in terms of discrimination. Does the item effectively separate based on the abilities of medical students in biochemistry? Third, to observe the detailed ability of students in responding to the items. A more detailed view of each item's ICC is presented in Fig. 2. Our findings indicate that there should be a spike (increase in ability) in student responses to items within the 1–3 ability range, as students in this range possess above-average abilities. However, in reality, students with high-order thinking skills did not respond effectively (unable to answer correctly), especially noticeable in items number 20, 8, and 46 (see the vertical red boxes). Their response abilities were relatively similar to those of students with low-order thinking skills. A notable increase was observed in items number 5, 7, 9, and 34 (see the vertical high green boxes). From this, we gather crucial information about which items require intervention and evaluation.

This section also highlights the need for internal consolidation in the study. This is a form of reflection for educators on the delivered material. It's important to avoid presenting items that are outside the core competencies formulated. Therefore, the role of the coordinator is crucial, as they must evaluate and consolidate immediately upon identifying any discrepancies.

The change in theta (θ) scores in biochemistry items indicates that the implementation of IRT is effective. This change can be observed in the ICC shown in Fig. 3. It's crucial to examine the relationship between student abilities and item difficulty levels within the framework of latent variables, a possibility afforded by the mathematical model system in IRT [10]. This IRT technique also aims to enhance accuracy in performance classification [11], [12].

A slight yet noticeable improvement in response ability was observed, as depicted in the ICC of the three revised items, numbers 20, 8, and 46 (see the short vertical green boxes). Although the increase is modest, the trend line shows an upward trajectory. Moreover, the ability performance no longer shows depression at the baseline. Item number 5 (Q5) also demonstrates an improvement in baseline ability (see the tall vertical green box) as illustrated in Fig. 4.

Key Findings

A significant observation in this study is that ICC can be used to analyze repeated test items. There's a common belief that reused items are of lower quality due to memorization by students. This is not entirely true or false, as the primary goal is to identify and select items that can effectively separate abilities. Utilizing ICC enhances precision in this selection. This study demonstrates that items like numbers 9 and 7 (Q9 and Q7) show clear ability separation despite being reused. ICC also aids in filtering out inconsistent items in terms of separation, such as item number 50 (Q50) (indicated by a long vertical red box). For item 50, it was observed that students with lower abilities had a high probability of guessing correctly. The ICC analysis also revealed transitions from difficult to easy items, as seen in items 39 and 44 (Q39 and Q44). Once high-quality items are identified, the next step is to create an item bank.

IRT can be adapted to produce more effective exam responses. A study by Ballis et al. found that exam frame design could enhance student responses in MCQ exams [13].

Conclusion

There was a significant change in average scores before and after IRT implementation, with the initial exam score being 56.1 and the second exam score 74.1 ($p < 0.01$). Item analysis revealed changes in theta scores for highly difficult items. An increase in baseline ability performance was observed in each item improved through the IRT approach, as indicated by the ICC curve patterns. These findings facilitate a more efficient and precise process of item identification and evaluation, particularly for biochemistry assessments in medical education.

Limitations

This study, while providing insightful findings, does have its limitations. The sample size, restricted to 89 medical students, may limit the generalizability of the results. Additionally, the study focuses exclusively on biochemistry assessments, which might not reflect the applicability of the findings to other subjects. Future research could benefit from a more diverse sample and a broader range of subjects to strengthen the generalizability of these findings.

Declarations

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

References

1. Z. Elhousni, J. LAAMECH, R. ZERHANE, and R. JANATI-IDRISSI, “Difficulties in learning biochemistry: Case of 1st year medical students, Tangier,” *Journal for Educators, Teachers and Trainers*, vol. 14, no. 1, pp. 63–74, Mar. 2023, doi: 10.47750/JETT.2023.14.01.006.
2. M. FMMT, W. KNH, L. SD, P. MYW, and P. PAJ, “Evaluation of the Teaching Approaches of Biochemistry for Medical Students: A Sri Lankan Case Study,” *Journal of Community Medicine & Health Education 2015 5:4*, vol. 5, no. 4, pp. 1–4, Aug. 2015, doi: 10.4172/2161-0711.1000359.
3. E. Wood, “Biochemistry is a difficult subject for both student and teacher,” *Biochem Educ*, vol. 18, no. 4, pp. 170–172, Oct. 1990, doi: 10.1016/0307-4412(90)90123-6.
4. Steven E. Stemler and Adam Naples, “Rasch Measurement v. Item Response Theory: Knowing When to Cross the Line,” *Practical Assessment, Research & Evaluation*, vol. 26, no. 11, pp. 1–6, 2021, doi: 10.7275/v2gd-4441.
5. P. F. M. Krabbe, “Item Response Theory,” *The Measurement of Health and Health Status*, pp. 171–195, Jan. 2017, doi: 10.1016/B978-0-12-801504-9.00010-6.
6. D. Zarate, B. A. Hobson, E. March, M. D. Griffiths, and V. Stavropoulos, “Psychometric properties of the Bergen Social Media Addiction Scale: An analysis using item response theory,” *Addictive Behaviors Reports*, vol. 17, p. 100473, Jun. 2023, doi: 10.1016/J.ABREP.2022.100473.
7. A. P. Kumar, A. Nayak, K. Manjula Shenoy, S. Goyal, and Chaitanya, “A novel approach to generate distractors for Multiple Choice Questions,” *Expert Syst Appl*, vol. 225, p. 120022, Sep. 2023, doi:

10.1016/J.ESWA.2023.120022.

8. M. Al-A'Ali, "IRT-Item Response Theory Assessment for an Adaptive Teaching Assessment System," in *10th WSEAS Interbational Conference on APPLIED MATHEMATICS*, Dallas: Applied Mathematics, 2006, pp. 1–5. Accessed: Oct. 10, 2023. [Online]. Available: https://www.researchgate.net/publication/240639436_IRT-Item_Response_Theory_Assessment_for_an_Adaptive_Teaching_Assessment_System
9. M. A. Dorner, P. Sadler, and B. Alters, "Still a private universe? Community college students' understanding of evolution," *Evolution: Education and Outreach*, vol. 16, no. 1, Dec. 2023, doi: 10.1186/S12052-022-00178-Y.
10. S. P. Reise and T. M. Moore, "Item Response Theory," *APA handbook of research methods in psychology: Foundations, planning, measures, and psychometrics (Vol. 1) (2nd ed.).*, pp. 809–835, Jun. 2023, doi: 10.1037/0000318-037.
11. A. E. Wyse and S. Hao, "An Evaluation of Item Response Theory Classification Accuracy and Consistency Indices," *Appl Psychol Meas*, vol. 36, no. 7, pp. 602–624, Oct. 2012, doi: 10.1177/0146621612451522.
12. L. M. Rudner, "Expected Classification Accuracy," *Practical Assessment, Research, and Evaluation*, vol. 10, no. 1, p. 13, Nov. 2019, doi: https://doi.org/10.7275/56a5-6b14.
13. B. Ballis, L. Lusher, and P. Martorell, "The effects of exam frames on student effort and performance," *Econ Educ Rev*, vol. 90, p. 102286, Oct. 2022, doi: 10.1016/J.ECONEDUREV.2022.102286.

Figures

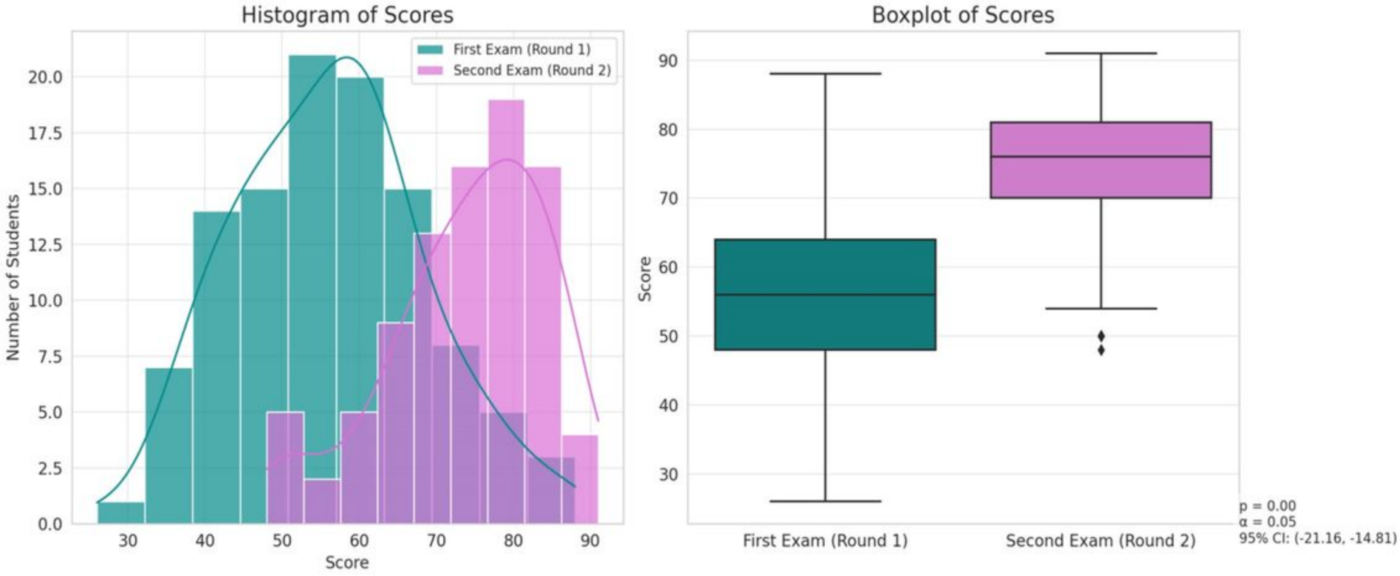

Figure 1

Wright Map illustrating the identification of difficult items in the assessment.

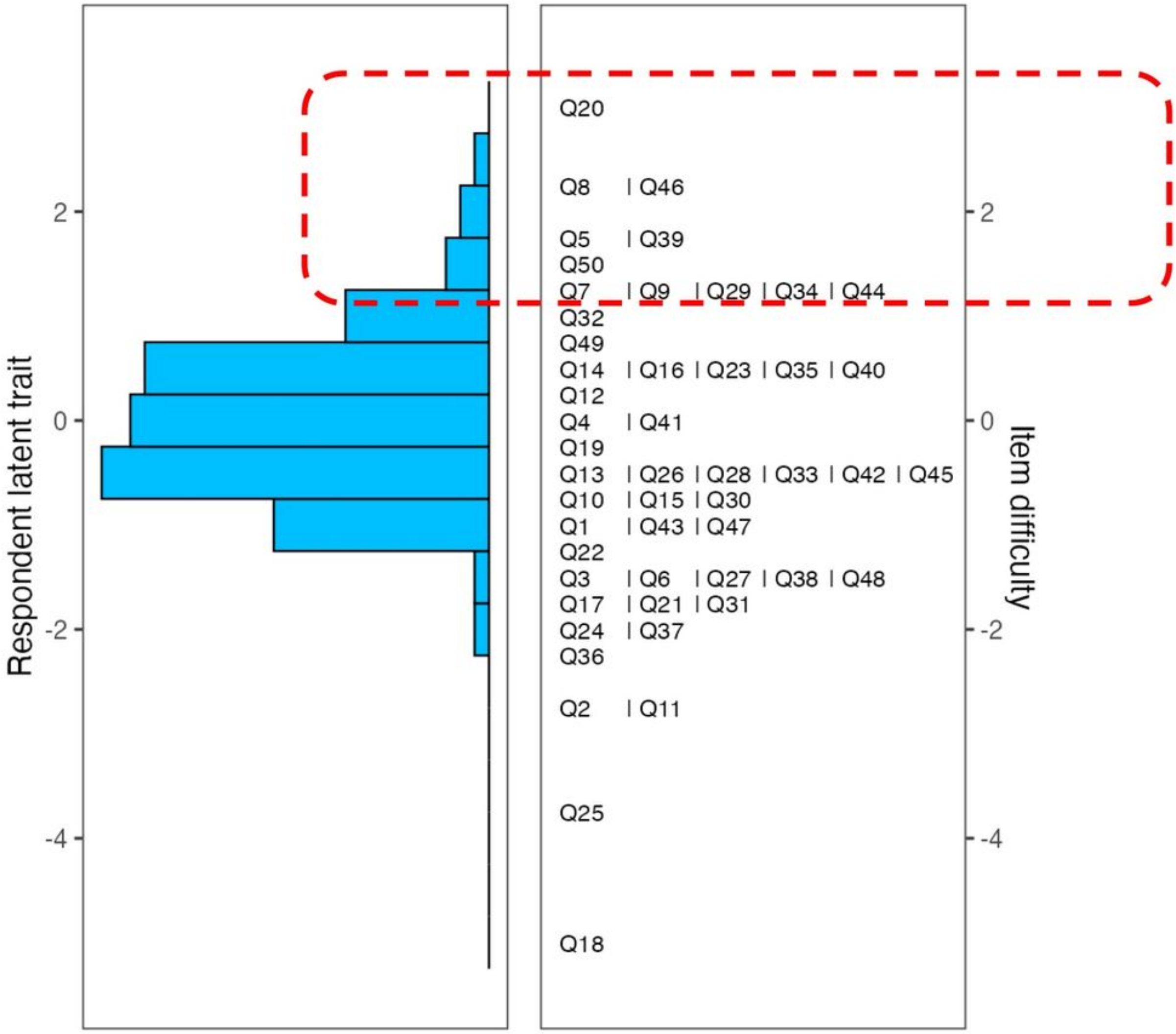

Figure 2

Wright Map to assist in the identification of difficult items in the assessment.

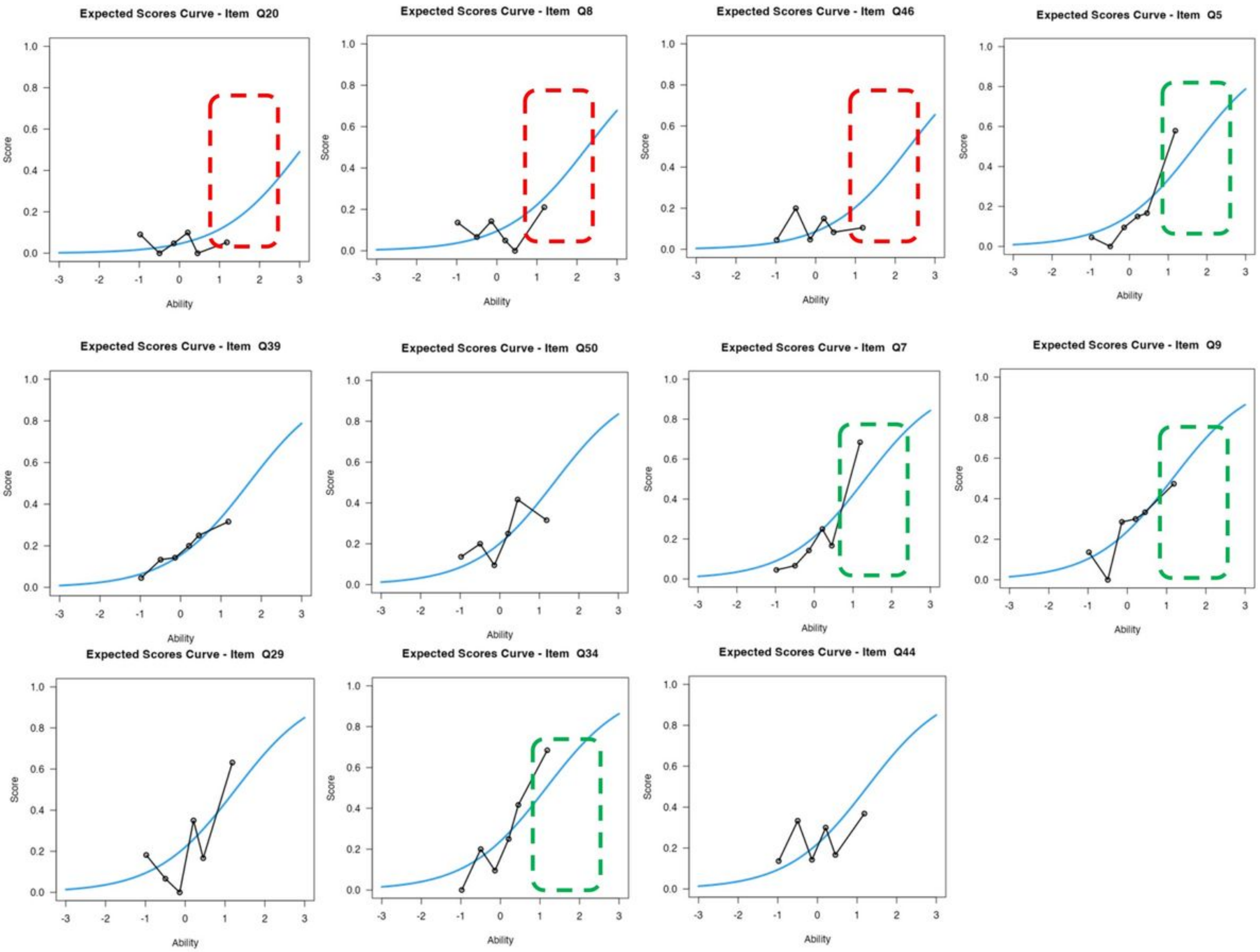

Figure 3

ICC of Items with High Difficulty Levels in the First Examination (Round 1).

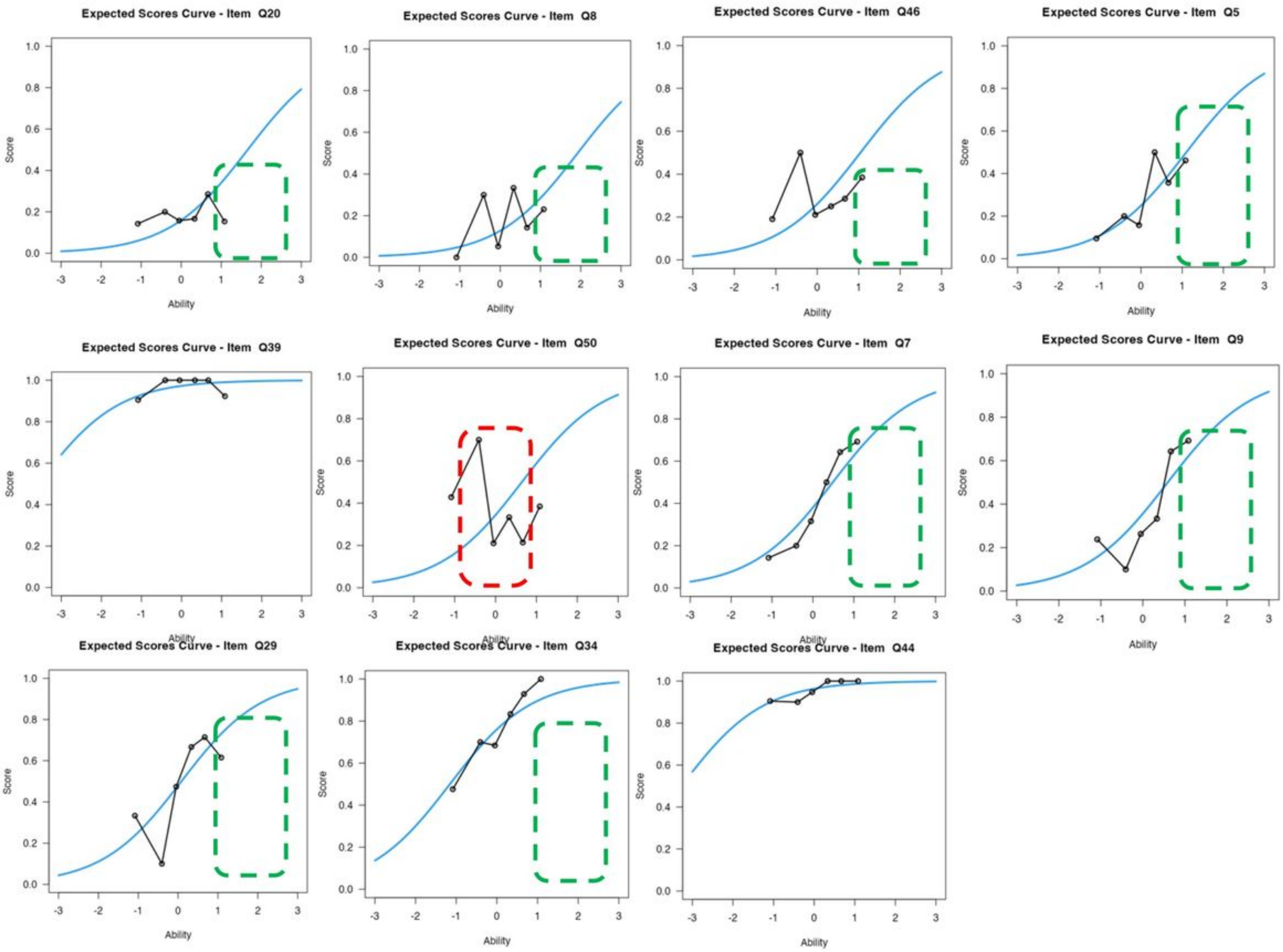

Figure 4

ICC for Items with High Difficulty Levels in the Second Examination (Round 2).