ASSESSING BIOLOGY ACHIEVEMENT USING THE RASCH MODEL

Boyxonov Islombek Abduqaxxor o‘g‘li *
Scientific-Study Practical Center under the Agency for Assessment Knowledge and Competences *

ANNOTATION: Biology achievement of 9 th-grade graduates from general secondary schools was assessed using the Rasch model. In this study, test results from 423 students were analyzed, and the difficulty levels of 48 test items and students’ ability levels were estimated on a common logit scale. Item fit to the Rasch model was examined using infit/outfit indices, and overall fit was found to be satisfactory. However, the outfit values of some items exceeded the acceptable threshold, indicating misfit to the model. Using the Wright map, the distribution of item difficulties and student abilities was examined, revealing that for certain ability ranges there were not enough items. Distractor analysis demonstrated the quality and depth of the items and confirmed that response options selected by fewer than 5 % of examinees were ineffective. Based on the findings, improving the biology test requires reviewing misfitting items, replacing weak distractors, and adding new items matched to different ability levels.

KEY WORDS: Rasch model; item difficulty; student ability; Wright map; infit and outfit; distractor analysis.

INTRODUCTION

The Rasch model plays an important role in improving the quality of test items and refining measurement processes. The Rasch model provides unidimensional, invariant (person-independent) measurement, supporting reliable and objective interpretation of test outcomes. Within this framework, person ability and item difficulty are expressed in the same logarithmic scale units (logits), which ensures linear measurement [3; 56-57 c.]. As a result, although direct comparisons of raw scores obtained from different test forms are limited, Rasch-calibrated test items allow scores to be compared within a common scale [4; 32-48 c.]. In assessing students’ biology achievement-especially for graduates of general secondary schools—using standardized tests makes it possible to fairly compare learning outcomes and differences in educational quality across schools and regions [1; 63-77 c.]. Therefore, in this study, biology test questions were analyzed in depth using the Rasch model, and the measurement properties of the test and item quality indicators were examined.

The purpose of the study was to evaluate 9 th-grade biology test items using the Rasch model and to determine the measurement quality of the test. Within this purpose, the following objectives were defined:

1. To calibrate item difficulty levels in the biology test using the Rasch model and to estimate students’ ability levels.
2. To evaluate the degree of model fit of item responses under the Rasch model (via infit/outfit statistics).
3. To compare items across multiple forms using the common-item test form method and to compare calibration results.
4. To analyze, using the Wright map, the distribution of item difficulties and student abilities in the test form and to determine the test’s coverage.
5. To develop recommendations for improving item quality and enhancing the measurement process based on the results.

For the analysis, biology test results from 423 9 th-grade graduates of general secondary schools were used. The test consisted of 48 questions, including 38 closed-format (“multiplikativ”) test items and 7 short-answer open-ended items. In addition, 1 item was presented in an extended open-ended format and was scored in 3 component parts (a total of 48 logically independent components/items) [2; 4 c.].

All items were dichotomously scored using a correct/incorrect principle; thus, the dichotomous version of the Rasch model was applied to each component. In the test form, each question had 3 or 4 answer options, with one correct answer (multiple-choice); each part of the extended item was scored separately.

Test outcomes were first reviewed using classical test theory (CTT) indices: for each item, the proportion correct (p-value) and discrimination (point-biserial correlation) were calculated. For example, for the B0000037-ID item-one of the most difficult-only 25 % of students answered correctly, and its discrimination index was below 0,25. A value below 0,25 indicates low discriminating power, and it is recommended that such an item be reviewed before inclusion in a test bank [3; 63-78]. For distractor analysis, selection percentages were computed for each response option, and incorrect options selected by fewer than 5% were treated as "non-functioning" distractors. For example, in test item 2, the selection rates were A=3 %, B=4 %, C=3 %, and D=89%, meaning two distractors (< 5 %) were almost never chosen [2; 6-7 c.]. Similar patterns were observed for other items, and a substantial share of the total 152 distractors operated below the 5% threshold.

At the next stage, results were analyzed in accordance with the Rasch model. For this purpose, the dexter package in the Python workflow and the R statistical environment were used. The dexter program estimates ability and difficulty parameters using the conditional maximum likelihood method (conditional maximum likelihood, CML) [4; 5 c.]. All responses from the 423 participants were assembled into a single matrix and jointly calibrated within one form.

For fit diagnostics, infit and outfit statistics were computed. Infit measures the extent to which observed responses for examinees around the item's targeted ability range match the model expectation, whereas outfit reflects unexpected deviations across the full sample based on the ε^2 criterion. In practice, outfit and infit values in the range 0,7-1,3 are commonly accepted as indicating adequate model fit. In this study, for each item, infit/outfit fit values and z-standardized (t) statistics were computed and interpreted using the specified criteria. In addition, a Wright map was produced from the Rasch results, placing the distribution of student abilities and item difficulties on the same scale. The Wright map can be used to identify where item difficulty is insufficient or excessive and to take corrective actions accordingly [5; 59-62 c.].

Rasch analysis of the 48-item biology test showed that, overall, the test results fit the model well. Infit statistics for all items fell within the specified acceptable interval, supporting the unidimensionality of the test. That is, within students' response patterns, no unexpected disordering (deviations from a Guttman scale pattern) was observed for any subgroup. However, outfit values for some items were high, indicating misfit to the model. In particular, the outfit index for item B0000037 exceeded the acceptable boundary. This suggests that for this item some students experienced unexpected difficulty or, conversely, answered correctly by guessing, meaning the item did not align with the overall response trend. The item characteristic curve (ICC) for B0000037 also showed substantial deviation from the model expectation: several points of the empirical function based on student responses fell outside the expected Rasch curve (even beyond the ±2 standard error bands). Therefore, this item was judged to be inappropriate for the test in terms of difficulty and should be improved before inclusion in a calibrated item bank.

Several other items-numbers 38, 40, 43, and 44 (ID: B0000031, B0000016, B0000043, B0000040)-also did not fully meet Rasch model requirements [3; 8-17 c.]. For these items, the distribution of student responses did not match the model at any ability group. In other words, students with very high or very low ability produced unexpected responses; thus, these items should also be reviewed before use in a calibrated test database.

Although the test's overall reliability (e.g., Cronbach's α or Rasch-scale reliability) was not reported in the articles, the fact that more than 85 % of the information obtained was attributable to student abilities in the –3 to +3 logit range indicates that the test results are reliable and measurement precision is sufficient [5; 59-80 c.].

Calibration yielded the difficulty parameter (β) for each test item. Estimated difficulties ranged from –2,63 to +2,41 logits [3; 65-70 c.]. Thus, the easiest item (β ≈ –2,6) was answered correctly by nearly all students, while the most difficult item (β ≈ +2,4) could be solved only by very strong students. The remaining items were of moderate difficulty, and the spread of difficulty values formed an almost continuous spectrum.

According to the Wright map, the distributions of item difficulty and student ability were aligned in a largely appropriate manner. Within the –3 to +3 logit ability interval, the test can provide strong differentiation and measurement precision; in this range, more than 85% of the total information was obtained. This indicates that for the majority of students (approximately 85%), the test is well targeted to their ability level and provides reliable assessment. However, a more detailed Wright map inspection showed that for some students with very high ability (e.g., above +3 logits) there were not enough difficult items, and conversely, for those with very low ability (below –3 logits) there were not enough very easy items [5; 62-70 c.]. For example, relatively few items were available for examinees above +2,5 logits and below –2,5 logits. This implies reduced test informativeness for assessing achievement at these extreme ability ranges.

Although the biology test was administered in only one form, the issue of comparing outcomes across different forms was also considered. Theoretically, if multiple forms of the same test are administered to different groups and some items overlap across forms (i.e., an item from one test also appears in another), the Rasch model can link (equate) these forms on a single measurement scale. Within this study, such a situation was artificially modeled, and results from 4 conditional test forms were analyzed. Each form was linked by 6 common items (identical questions). As a result, Rasch calibration showed that the difficulty levels of the common items were nearly identical across all forms. This indicates that the different test forms were placed on the same measurement scale.

At the same time, it was shown that comparing raw scores may not yield valid conclusions. For example, although a student who answered 20 out of 30 moderately difficult items correctly in one form and a student who answered 20 out of 30 much easier items correctly in another form have the same raw score (20), their Rasch ability estimates differ [4; 32-40 c.]. Therefore, directly comparing uncalibrated raw scores across standard test forms can lead to incorrect conclusions, whereas the Rasch model enables fair comparisons by linking forms through common items.

Overall, Rasch-based calibration of the 48 biology test items supports scientifically grounded identification of the test's measurement properties and contributes to standardization. The broad difficulty range (–2,6 to +2,4 logits) and its alignment with student abilities indicate that the test can be reused in subsequent assessments. Fit diagnostics suggest that the test predominantly measures a unidimensional construct; however, high outfit values for certain items (e.g., B0000037) indicate a need to improve those items or exclude them from operational use. Wright map evidence also highlights insufficient item coverage at very high and very low ability levels, implying that the test should be enriched with both more difficult and easier items to expand coverage. Distractor analysis likewise indicates that response options selected by fewer than 5 % function poorly and should be revised or replaced.

In conclusion, calibrating the 48 biology test items using the Rasch model demonstrates that the test's measurement characteristics can be established on a scientific basis and that the test can be standardized. Item difficulties spanned a wide range from –2,6 to +2,4 logits and were distributed in accordance with student ability levels, enabling reuse of the test in future assessments. Model-fit diagnostics confirmed that the test largely measures a unidimensional construct; however, elevated outfit for some items (e.g., B0000037) suggests that these items require improvement or removal. Wright map analysis revealed a shortage of items at very high and very low ability ranges, indicating that the test should be expanded with both more complex and simpler items to improve coverage. Establishing a calibrated item bank, implementing common-item linking and equating across forms, and introducing these procedures can enable fair comparisons across schools and regions and increase the validity and reliability of assessment in national certification examinations and educational quality monitoring. In addition, to further strengthen fairness, conducting Differential Item Functioning (DIF) analyses and checking item neutrality across different groups is an important direction for future research.

REFERECES:

[1] A.B. Normurodov, M.Dj. Ermamatov, A.A. Baratov, I.A. Boyxonov. Umumiy o‘rta ta’lim maktablarining 9–sinf bitiruvchilari uchun biologiya fanidan bilimlarni baholashda standart testlardan foydalanish, BMBA “Axborotnoma” ilmiy-uslubiy jurnali, 2023-yil, 1-son, 63–77-betlar.

[2] M.Dj. Ermamatov, I.A. Boyxonov. Distraktorlar tahlili: biologiya fanidan o‘tkazilgan test sinovi natijalari, BMBA “Axborotnoma” ilmiy-uslubiy jurnali, 2023-yil, 2-son, 4–19-betlar.

[3] A.B. Normurodov, M.Dj. Ermamatov, I.A. Boyxonov. Rash modeli bilan moslik: biologiya fanidan o‘tkazilgan test sinovi natijalari, BMBA “Axborotnoma” ilmiy-uslubiy jurnali, 2023-yil, 3-son, 56–83-betlar.

[4] M.Dj. Ermamatov, I.A. Boyxonov. Umumiy elementli test variantlar: biologiya fanidan olingan test natijalari tahlili, BMBA “Axborotnoma” ilmiy-uslubiy jurnali, 2023-yil, 2-son, 51–72-betlar.

[5] A.B. Normurodov, I.A. Boyxonov. Rayt xaritasi, ichki va tashqi moslik statistikasi: biologiya fanidan o‘tkazilgan test sinovi natijalari, BMBA “Axborotnoma” ilmiy-uslubiy jurnali, 2023-yil, 2-son, 73–96-betlar.