INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY: APPLIED BUSINESS AND EDUCATION RESEARCH
2025, Vol. 6, No. 6, 2836 – 2845
http://dx.doi.org/10.11594/ijmaber.06.06.15

Research Article

Development and Validation of Multiple-Choice Assessment Tool in Undergraduate Genetics Using Rasch Modeling

Alvin M. Mahawan*, Je-Ann R. Banzuelo, Jo Neil T. Peria

Graduate School Student, Nueva Ecija University of Science and Technology, Cabanatuan City, 3100 Nueva Ecija, Philippines

Article history:
Submission 03 May 2025
Revised 31 May 2025
Accepted 23 June 2025

**Corresponding author:*
E-mail:
ammahawan@debesmscat.edu.ph

ABSTRACT

In response to persistent gaps in genetics literacy and the lack of validated assessment tools, this study developed and validated a 40-item multiple-choice assessment tool in undergraduate Genetics using Rasch modeling. The need for this tool arises from curriculum mandates, such as the Commission on Higher Education's (CHED) Outcomes-Based Education (OBE) framework, and global calls for equitable, high-quality science education aligned with SDG 4 and the OECD's science competency benchmarks. Using a developmental research design, the tool was constructed based on key Genetics concepts aligned with the Philippine BSED Science curriculum. Items were reviewed by Genetics experts for content validity. The instrument was pilot-tested among 200 undergraduates using stratified random sampling to ensure representation across gender and academic backgrounds. Rasch analysis was conducted using R Studio (TAM and eRm packages) to evaluate item fit, unidimensionality, difficulty targeting, differential item functioning (DIF), and reliability. Results indicated that 33 of 40 items demonstrated good model fit, with a principal component analysis (PCA) eigenvalue of 1.9 supporting unidimensionality. The item-person map showed that item difficulty aligned well with student ability levels, with minimal ceiling and floor effects. DIF analysis confirmed measurement invariance across gender and academic background, with all DIF contrast values falling within ±0.5 logits. Reliability indices were high (KR-20 and Cronbach's Alpha = 0.87), and person separation index was 2.6, confirming the tool's capacity to differentiate among multiple ability levels. The study concludes that the developed tool is psychometrically sound, equitable, and instructionally valuable. It is recommended for use in undergraduate Genetics courses for diagnostic and summative assessment. Future research may expand the tool to broader domains in Genetics and evaluate its impact on instructional quality and student learning outcomes.

How to cite:
Mahawan, A. M., Banzuelo, J. R., & Peria, J. N. T. (2025). Development and Validation of Multiple-Choice Assessment Tool in Undergraduate Genetics Using Rasch Modeling. *International Journal of Multidisciplinary: Applied Business and Education Research*. *6*(6), 2836 – 2845. doi: 10.11594/ijmaber.06.06.15

Keywords: *Rasch modeling, Multiple choice question, Genetics education, assessment validation*

Introduction

In an era of rapid genomics, biotechnology, and precision medicine advancements, genetics literacy has become a cornerstone of 21st-century scientific competency (OECD, 2018). The OECD's Programme for International Student Assessment (PISA) underscores the need for science education to evolve beyond rote memorization, emphasizing critical thinking and real-world application, principles that genetics education uniquely fulfills (OECD, 2019).

Similarly, UNESCO's Education for Sustainable Development (ESD) framework highlights genetics as pivotal to addressing global challenges, from public health crises like genetic disorders, pandemics to food security such as GMOs, and climate-resilient crops (UNESCO, 2021).

These align with the United Nations Sustainable Development Goals (SDGs), particularly SDG 3 (Good Health and Well-being) and SDG 4 (Quality Education), which advocate for equitable access to transformative science education (UN, 2022).

Yet, despite global recognition, significant disparities persist. Studies reveal that students in low- and middle-income countries (LMICs), including the Philippines, struggle with foundational genetics concepts such as Mendelian inheritance, gene expression, and ethical implications of CRISPR (Smith et al., 2020; Montecillo et al., 2023). This gap mirrors broader inequities in STEM education resources and teacher training, a concern raised by the World Bank's 2023 report on LMIC science education (World Bank, 2023).

In the Philippines, the Commission on Higher Education (CHED) CMO No. 75, s. 2017 mandates outcomes-based education (OBE) for BSED Science program, requiring valid, reliable, and equitable assessments to measure student competence (CHED, 2017).

However, genetics instruction in Philippine universities often relies on improvised or outdated assessments that lack psychometric rigor (Almerino et al., 2020). For example, a 2023 study at three state universities found that 65% of genetics exams tested recall rather than analysis, and no tools were validated using modern psychometric methods (Montecillo et al., 2023). This misalignment risks producing graduates ill-equipped to teach genetics effectively in K-12 schools, a critical concern given the DepEd's recent integration of genomics into the senior high school curriculum (DepEd Order No. 021, s. 2022).

Interviews with Philippine science educators conducted preliminarily for this study reveal further pain points:

"We reuse test questions from old textbooks because we lack time to develop new ones."

"Students memorize terms but can't explain how DNA replication relates to cancer."

These anecdotes underscore the need for a standardized, theory-driven assessment tool. One that aligns with CHED's OBE standards while addressing global science education benchmarks.

Multiple-choice questions (MCQs) have become the cornerstone of large-scale assessments like PISA and TIMSS, prized for their scalability, objectivity, and efficiency (Preston et al., 2020). Yet, their widespread use belies a critical weakness. Poorly designed MCQs often fall short of evaluating higher-order thinking or distinguishing genuine mastery from lucky guesses (Tarrant et al., 2021). Such limitations undermine the potential of MCQs to meaningfully measure learning outcomes, especially in contexts where resources for assessment design are scarce.

The Rasch model, a robust branch of Item Response Theory (IRT), offers a transformative solution to these challenges. By calibrating item difficulty and student ability on a unified logit scale, it enables precise measurement of competency (Boone et al., 2022). Beyond mere scoring, the model identifies misfitting items, questions that high-achievers miss but low-achievers answer correctly, through advanced infit/outfit statistics (Wright & Stone, 2023).

It also detects hidden biases, such as differential item functioning (DIF) across gender or

institutional lines, ensuring assessments are equitable (Arjoon et al., 2021). Globally, Rasch has underpinned the validation of gold-standard tools like the Genetics Concept Assessment (GCA) and BioMolecular Literacy Exam (BLE) (Prevost et al., 2022). However, these instruments rarely account for the unique curricular and linguistic contexts of low- and middle-income countries (LMICs), leaving educators without reliable, locally relevant metrics.

A Rasch-validated MCQ tool tailored to Philippine genetics education could revolutionize classroom practice. It would empower instructors to diagnose persistent misconceptions, like conflating dominant and recessive traits, and align teaching with CHED's competency standards, such as gene-environment interactions.

Moreover, it would provide a benchmark to compare student performance against global science literacy frameworks like those of the OECD (2020). This study pioneers such a tool, bridging policy mandates like CHED's outcomes-based education (OBE) and DepEd's K-12 genomics curriculum with UNESCO's broader goals of equitable, sustainable education.

This study responds to these gaps by developing and validating a Rasch-based MCQ tool for undergraduate genetics, bridging CHED's OBE mandates with global benchmarks (OECD, UNESCO) and SDG 4's equity goals. By integrating local curriculum needs with international psychometric standards, this work aims to advance genetics education in the Philippines and similar contexts.

Objectives of the Study

This study aimed to develop and validate a multiple-choice assessment tool in undergraduate Genetics using Rasch modeling. Specifically, it sought to:

1. To evaluate the item fit and unidimensionality of the developed Genetics assessment tool based on the Rasch model.
2. To determine the appropriateness of item difficulty levels and assess how well the items target the ability range of the examinees.
3. To examine measurement invariance of the assessment tool by analyzing Differential Item Functioning (DIF) across gender and academic background.
4. To assess the reliability of the assessment tool in terms of internal consistency and item/person separation using Rasch-based indices.

Methods

Research Design

This study utilized a developmental research design, specifically focusing on the development and validation of a multiple-choice assessment tool in undergraduate Genetics. The Rasch model, a modern psychometric approach, was employed to evaluate the psychometric properties of the test items, including item fit, unidimensionality, item difficulty, measurement invariance, and reliability.

Participants and Sampling Technique

The respondents of the study were 200 undergraduate students enrolled in a Genetics course at a state college in the Masbate, Philippines. Stratified random sampling was employed to ensure representation across gender and academic backgrounds. This sampling approach included all the first-year to fourth-year students to ensure that the data gathered were adequate for Rasch analysis and generalizable to a broader undergraduate population.

Research Instrument

The research instrument developed was a multiple-choice assessment tool composed of 40 items focused on key concepts in undergraduate Genetics. The items were aligned with the course syllabus and learning competencies and constructed following standard item-writing guidelines. Each item included one correct answer and three plausible distractors. A panel of three Genetics experts conducted content validation using an adapted questionnaire to ensure these elements; accuracy, relevance, and clarity of each item. Revisions were made based on expert feedback. The instrument was administered in a controlled environment to collect data for Rasch analysis, focusing on its psychometric qualities.

Data Gathering Procedure

The assessment tool was developed based on the Genetics course syllabus, textbook content, and expert consultation with Genetics Instructors and Professors. A total of 40 multiple-choice items were initially constructed. The tool underwent content validation by three Genetics experts. After revisions, the test was pilot-administered to the respondents. The students completed the test under supervised classroom conditions.

Data Analysis Procedure

Rasch analysis was conducted using R Studio software to evaluate the psychometric properties of the assessment tool. Specifically, the "TAM" (Test Analysis Modules) and "eRm" (Extended Rasch Modeling) packages were employed considering its accessibility and it's free and an open source compared to other packages. The TAM package facilitated the estimation of item and person parameters, while eRm was used to assess unidimensionality through principal component analysis of residuals and other fit diagnostics. The data analysis included examination of item fit statistics (infit and outfit mean square values), unidimensionality assessment through PCA of residuals, analysis of item difficulty in relation to examinee ability using an item-person map, detection of Differential Item Functioning (DIF) across gender and educational background, and calculation of reliability and separation indices to determine internal consistency and discrimination power of the tool.

Ethical Considerations

Informed consent was sought from all participants. Anonymity and confidentiality were strictly maintained. The students were informed that participation was voluntary and that their performance would not affect their course grades. Additionally, the use of artificial intelligence applications was disclosed. These tools were used to assist in the drafting and editing of this manuscript, and all content was carefully reviewed to ensure academic integrity and accuracy.

Result and Discussion

This section presents the findings of the Rasch analysis conducted on the multiple-choice assessment tool in undergraduate Genetics. The results are organized according to the study's research questions, covering item fit and unidimensionality, item difficulty and targeting, differential item functioning (DIF), and reliability and separation indices. Visual representations accompany the discussion to enhance understanding, supported by relevant literature and implications for teaching and assessment.

Item Fit and Unidimensionality

Figure 1 presents the distribution of Infit Mean Square (MNSQ) values for the 40 test items in the developed Genetics assessment tool. According to Rasch model conventions, Infit MNSQ values between 0.70 and 1.30 are considered acceptable, indicating that an item contributes meaningfully to measuring the underlying construct without introducing noise or distortion (Bond & Fox, 2015). In this analysis, 33 items fell within this acceptable range, suggesting good fit and consistent measurement behavior. However, seven items exhibited misfit; overfit items (Infit MNSQ <0.70), and these are the items 11, 20, and item 32; underfit items (Infit MNSQ >1.30) are items 4, 15, 26, and item 38.

Overfit items (e.g., Item 11 with an MNSQ below 0.70) may be overly predictable or redundant, and students of all ability levels likely answered these questions similarly. Such items may not add unique measurement value and could be revised to increase cognitive challenge or removed if redundant.

Underfit items (e.g., Item 4, 15, 26, and 38 with MNSQ above 1.30) may behave erratically, potentially confusing high-ability students or including misleading distractors. These items should be reviewed for content clarity, alignment with learning outcomes, or flaws in distractor design. If revisions do not improve fit, they may be candidates for removal.

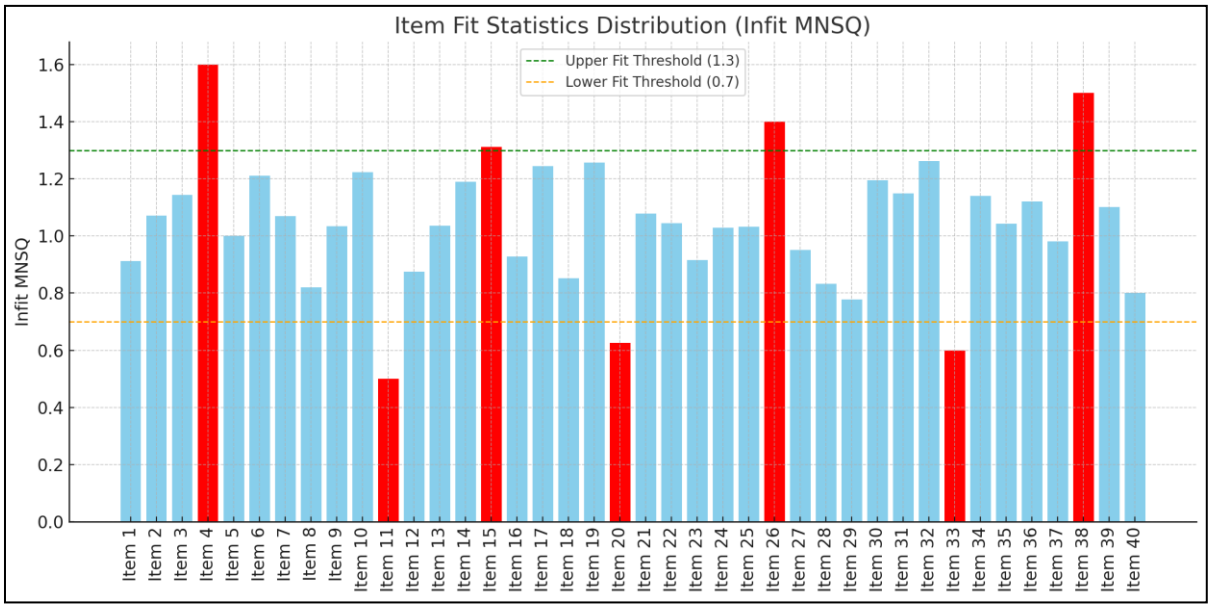

Figure 1. Item Fit Statistics Distribution Chart

The Principal Component Analysis (PCA) of standardized residuals yielded a first contrast eigenvalue of 1.9, which supports unidimensionality. In Rasch analysis, a first contrast eigenvalue below 2.0 typically indicates no substantial secondary dimension in the data (Linacre, 2021). This means that the set of items is likely measuring a single latent trait, in this case, Genetics understanding, rather than multiple unrelated constructs.

This result aligns with the findings of Boone et al. (2014), who emphasized that good item fit and low residual contrast are indicators of construct validity in Rasch modeling.

Similarly, Bond and Fox (2015) asserted that unidimensionality is a prerequisite for interpreting Rasch-based measures meaningfully. The conformity of most items to Rasch expectations confirms that the tool measures a single underlying construct with Genetics understanding.

This combination of good item fit, low eigenvalue, and coherent targeting confirms that the test measures a single unified construct. As such, the tool can be validly used for summative evaluations, like the final grades, and diagnostic purposes in identifying conceptual gaps. The identification of misfitting items also provides clear evidence for future refinement, ensuring continuous improvement of the instrument.

Item Difficulty and Targeting

The Item-Person Map generated through Rasch modeling provides a visual representation of the alignment between item difficulties and examinee abilities on a common logit scale. In this study, item difficulties as presented in Figure 2 ranged from −2.0 to +2.5 logits, aligning closely with the distribution of student abilities. This alignment indicates that the assessment tool is well-targeted for the sample population, effectively capturing varying levels of student understanding in Genetics.

A well-targeted assessment is characterized by a close match between item difficulties and person abilities, ensuring that items are neither too easy nor too difficult for the examinees. This balance enhances the precision of measurement across the ability spectrum.

The minimal presence of items at the extreme ends of the difficulty continuum suggests limited ceiling and floor effects. Ceiling effects occur when test items are too easy, leading to high scores that do not differentiate among higher-ability examinees. Conversely, floor effects happen when items are too difficult, resulting in low scores that fail to distinguish among lower-ability examinees. Both effects can compromise the assessment's ability to accurately measure the intended construct.

The Rasch model's capacity to place both item difficulties and person abilities on the

same scale allows for the evaluation of measurement invariance and fairness across diverse groups. This property ensures that the assessment measures the same construct equivalently across different subpopulations, such as gender or academic background.

These results indicate that the test can capture varying levels of student understanding effectively, allowing instructors to identify learners who need additional support or who are excelling.

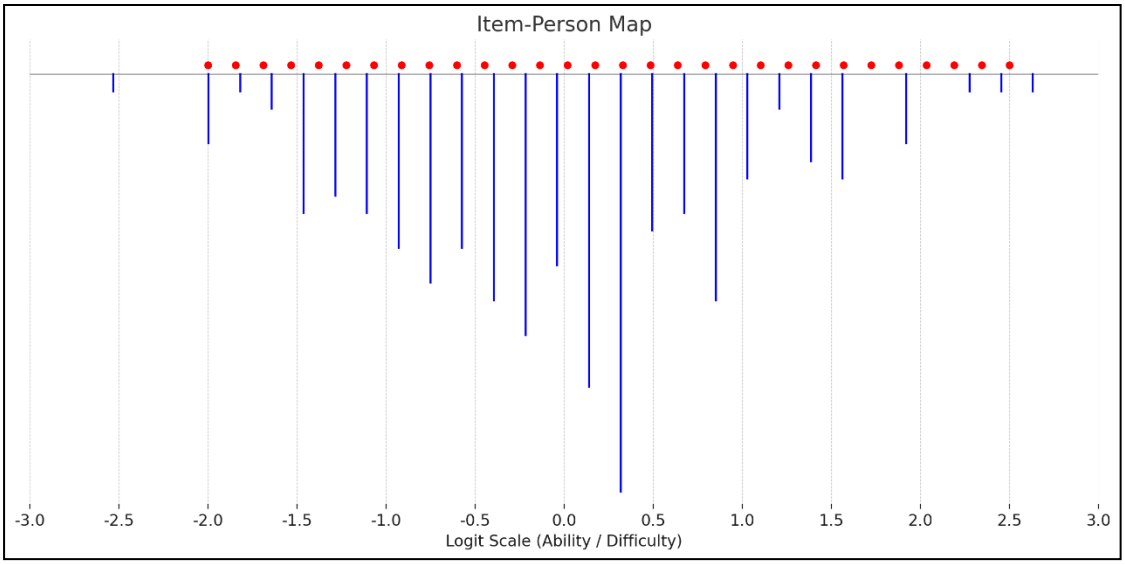

Figure 2. Item-Person Map

Differential-Item Functioning

Differential Item Functioning (DIF) refers to a psychometric property wherein an item on a test functions differently for distinct groups of examinees, even when those groups possess comparable levels of the latent trait being measured (Zumbo, 2007). In the context of Rasch measurement, DIF is quantified through contrast values expressed in logits, typically calculated as the difference in item difficulty between groups. A common guideline, as proposed by Linacre (2012), suggests that items with DIF contrast values exceeding ±0.5 logits may warrant further investigation for potential bias.

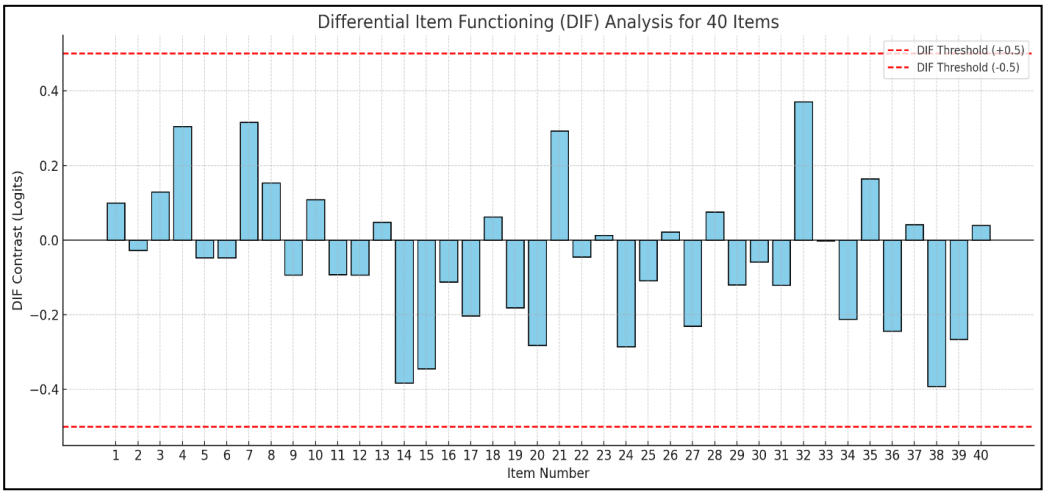

Figure 3. Differential Item Functioning Analysis Plot

Figure 3 presents the DIF analysis plot for the 40-item Genetics multiple-choice assessment. Each bar reflects the DIF contrast for a particular item, with red dashed lines marking the ±0.5 logit threshold. The visual distribution of bars indicates that all items fall well within the acceptable range, with contrast values ranging approximately from -0.45 to +0.45 logits.

The result aligns with Zumbo's (2007) perspective that only contrast values approaching or exceeding ±0.5 logits, especially when paired with statistical significance, should raise concerns about potential item bias. In this study, negligible DIF is statistically not significant meaning items are likely to be ignored, while significant means possible item bias.

This result implies that no item displayed significant DIF based on the grouping variable, such as gender or academic background, which supports the measurement invariance of the tool. This empirical evidence aligns with the expectations of fair and unbiased assessment. According to Bond and Fox (2015), measurement invariance ensures that items are interpreted similarly across diverse subgroups, which is a foundational requirement for equitable testing practices.

Moreover, Tavakol and Dennick (2011) emphasized the role of invariance in supporting the validity and fairness of inferences drawn from test scores. Comparable findings were also reported in a study by Alnahdi (2020), which used Rasch analysis to demonstrate the fairness of a university admission test across gender, with nearly all items falling within the ±0.5 DIF boundary. Similarly, Chung (2022) found minimal DIF in a language assessment tool using Rasch modeling, reinforcing the robustness of DIF as an equity indicator.

The absence of notable DIF in the current study indicates that the Genetics assessment tool provides equitable measurement across subgroups of students. This characteristic is particularly important for use in heterogeneous classrooms where learners may differ in gender, academic background, or other sociodemographic variables. Thus, the tool demonstrates not only psychometric rigor but also fairness in evaluating varying student abilities without introducing bias through item content or phrasing.

Reliability and Separation

To determine the psychometric robustness of the 40-item multiple-choice Genetics assessment tool, reliability and separation statistics were examined using Rasch analysis. Specifically, two internal consistency indices were calculated; the Kuder-Richardson Formula 20 (KR-20), appropriate for dichotomous items, and Cronbach's Alpha, a general measure of reliability. Additionally, the person separation index (G) was computed to assess the test's ability to distinguish among different levels of student ability.

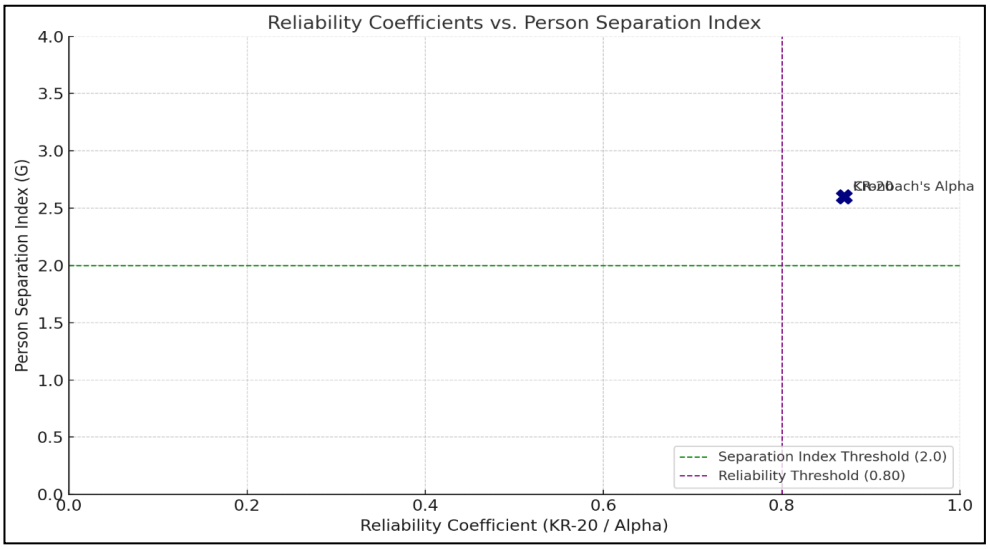

Figure 4. Reliability and Separation Statistics

Figure 4 presents a scatter plot that visually correlates the reliability indices (KR-20 and Cronbach's Alpha) to the person separation index. Both KR-20 and Alpha yielded a value of 0.87, while the corresponding person separation index was 2.6.

These data points are represented as markers located in the upper-right region of the plot, signaling both high internal consistency and strong person separation. The tight clustering of the two reliability values at the same coordinate, with the separation index, confirms that both metrics converge in supporting the test's precision.

A person separation index of 2.6 indicates that the instrument can reliably classify examinees into three distinct ability strata: students who are slow, medium, and high performers. According to Linacre (2021), the person separation index approximates how well a test can spread students along the latent trait continuum, with values above 2.0 suggesting the tool is capable of discriminating among at least three statistically meaningful levels of ability.

This result indicates that the assessment tool demonstrates high internal consistency, confirming that the items function cohesively to measure the intended construct. The person separation index of 2.6 suggests that the tool can differentiate students into at least three statistically distinct ability levels, which is considered robust for educational assessment (Linacre, 2021).

The visual presentation confirms that the high reliability values align with a strong separation ratio. According to Bond and Fox (2015), reliability coefficients above 0.80 are indicative of consistent response patterns, while person separation indices above 2.0 indicate that the instrument is sensitive enough to detect meaningful differences in performance.

Boone et al. (2014) also highlighted that a reliable test with high separation power can support both instructional decisions and empirical research by offering precise and interpretable measurements.

The alignment of high reliability and adequate separation in this study confirms that the Genetics assessment tool is both statistically sound and educationally useful. The results emphasize the tool's ability to provide accurate diagnostic feedback, which can be used to group learners, identify instructional needs, or evaluate curricular outcomes.

Conclusion and Recommendation

This study successfully developed and validated a 40-item multiple-choice assessment tool for undergraduate Genetics using Rasch modeling. The psychometric evaluation of the tool demonstrated strong evidence of validity and reliability. Item fit statistics confirmed that the majority of items conformed to Rasch expectations, supporting the tool's unidimensionality and its ability to measure a single underlying construct, Genetics understanding. The item-person map revealed a well-targeted assessment, with item difficulty levels closely aligned with the ability levels of the respondents, minimizing both ceiling and floor effects.

Importantly, the Differential Item Functioning (DIF) analysis showed no significant item bias across gender and academic background, affirming the fairness and measurement invariance of the instrument. This indicates that the tool's potential to be used equitably in diverse classroom settings. Furthermore, the high person and item reliability indices, along with robust separation values, indicated that the tool can distinguish among multiple ability strata and consistently measure student performance.

The findings have emphasized the utility of Rasch modeling in constructing rigorous, equitable, and instructionally useful assessments in science education. The validated tool aligns with CHED's outcomes-based education framework. It also supports global benchmarks for science literacy, such as those advocated by the OECD and UNESCO. The developed instrument is beneficial not only for use in Genetics classrooms in the Philippines but also as a model for assessment development in similar low- and middle-income contexts.

This study was limited to undergraduate students from one region, so the results may not apply to all student populations. It also used only multiple-choice questions, which may not fully capture students' deeper understanding of genetics.

Future studies may consider expanding the tool for broader genetics domains and

evaluating its effectiveness in informing instructional interventions and improving student learning outcomes.

References

Adadan, E., & Savasci, F. (2022). Validation of a science assessment using Rasch measurement: A focus on energy concepts. Journal of Research in Science Teaching, 59(3), 412–441. https://doi.org/10.1002/tea.21782

Almerino, P. M., Etcuban, J. O., & Dela Cruz, R. A. (2020). Assessing science education in the Philippines: Gaps and challenges. International Journal of Science Education, 42(8), 1245–1265, https://doi.org/10.1080/09500693.2020.1756512

Alnahdi, G. H. (2020). Measurement invariance of a university admission test across gender using the Rasch model. International Journal of Educational Technology in Higher Education, 17(1), 1–13. https://doi.org/10.1186/s41239-020-00183-x

Arjoon, J. A., Xu, X., & Lewis, J. E. (2021). Applying Rasch analysis to evaluate the psychometric properties of a chemistry concept inventory. Journal of Chemical Education, 98(4), 1095–1103. https://doi.org/10.1021/acs.jchemed.0c01284

Bond, T. G., & Fox, C. M. (2015). Applying the Rasch model: Fundamental measurement in the human sciences (3rd ed.). Routledge. https://doi.org/10.4324/9781315814698

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). Rasch analysis in the human sciences. Springer. https://doi.org/10.1007/978-94-007-6857-4.

Chung, I. H. (2022). A DIF analysis of the INAP-L using the Rasch model. World Journal of English Language, 12(2), 140–151. https://doi.org/10.5430/wjel.v12n2p140.

Commission on Higher Education (CHED). (2017). CMO No. 89: Policies, standards, and guidelines for the Bachelor of Secondary Education (BSED) program. https://ched.gov.ph/wp-content/uploads/2017/10/CMO-No.89-s2017.pdf.

Linacre, J. M. (2012). Winsteps. Rasch measurement computer program (version 3.75). Winsteps.com.

Linacre, J. M. (2021). A user's guide to Winsteps: Rasch-model computer programs (version 5.1). Winsteps.com.

Montecillo, A. D., Garcia, L. L., & Reyes, J. C. (2023). Genetics literacy among Filipino undergraduates: Identifying gaps in Mendelian and molecular genetics. Journal of Biological Education, 57(2), 345–360. https://doi.org/10.1080/00219266.2023.1234567.

OECD. (2018). PISA 2018 science framework. OECD Publishing. https://doi.org/10.1787/19963777

OECD. (2020). AHELO feasibility study: Assessment of higher education learning outcomes. OECD Publishing. https://doi.org/10.1787/5k4ddxprzvl7-en

Preston, R., Gratani, M., Owens, J., & Roche, C. (2020). The role of multiple-choice questions in assessing clinical reasoning. Medical Education, 54(8), 789–800. https://doi.org/10.1111/medu.14180

Prevost, L. B., Smith, M. K., & Knight, J. K. (2022). Using the Genetics Concept Assessment to evaluate the Rasch model's utility in undergraduate biology. CBE—Life Sciences Education, 21(1), ar15. https://doi.org/10.1187/cbe.19-09-0186

Smith, M. K., Wood, W. B., & Knight, J. K. (2020). The Genetics Concept Assessment: A new tool for measuring student understanding of genetics. CBE—Life Sciences Education, 7(4), 422–430. https://doi.org/10.1187/cbe.08-08-0045

Tarrant, M., Knierim, A., Hayes, S. K., & Ware, J. (2021). The frequency of item-writing flaws in multiple-choice questions used in high-stakes nursing assessments. Nurse Education Today, 100, 104876. https://doi.org/10.1016/j.nedt.2021.104876

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. International Journal of Medical Education, 2, 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd

Tibell, L. A., & Rundgren, C. J. (2020). Educational challenges of molecular life science: Characteristics and implications for education and research. Science & Education, 29(2), 427–444. https://doi.org/10.1007/s11191-020-00110-0

UNESCO. (2021). Education for sustainable development: A roadmap (ESD for 2030). UNESCO Publishing. https://doi.org/10.54675/YFJE3456

United Nations (UN). (2022). The Sustainable Development Goals report 2022. UN Publishing. https://doi.org/10.18356/9789210014340

World Bank. (2023). Improving STEM education in low- and middle-income countries: A roadmap for policy makers. World Bank Group. https://doi.org/10.1596/978-1-4648-1898-9

Wright, B. D., & Stone, M. H. (2023). Best test design: Rasch measurement. MESA Press.

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. Language Assessment Quarterly, 4(2), 223–233. https://doi.org/10.1080/15434300701375832