MODESTUM EURASIA Journal of Mathematics, Science and Technology Education, 2026, 22(3), em2790
ISSN:1305-8223 (online)
OPEN ACCESS Research Paper https://doi.org/10.29333/ejmste/18065

From midterm to final: Using Rasch analysis to measure growth and fairness in university calculus assessments

Shirali Kadyrov [1,2*], Rosemary Callingham [3], Alfira Makhmutova [1], Goncalo Pinto [1]

[1] Department of General Education, New Uzbekistan University, Tashkent, UZBEKISTAN
[2] Department of Mathematics and Natural Sciences, SDU University, Kaskelen, KAZAKHSTAN
[3] School of Education, University of Tasmania, Hobart, AUSTRALIA

Received 03 June 2025 ▪ Accepted 29 January 2026

Abstract

This study investigates the reliability, validity, and fairness of university-level calculus assessments using Rasch analysis, focusing on the progression and alignment of key concepts across midterm and final exams. A cohort of 369 students from diverse academic disciplines was assessed using standardized exams designed to align with learning outcomes. The midterm and final exams, comprising conceptual and computational items, were analyzed separately and combined to evaluate their fit to the Rasch model. Summary statistics confirmed the reliability of item and person measures, while alignment analysis highlighted conceptual continuity across assessments. The results demonstrated the assessments' efficacy in consistently measuring student ability across different subgroups and performance levels. Wright maps and item-level statistics provided a comprehensive view of student understanding, identifying areas requiring targeted pedagogical intervention. The study revealed significant growth in student performance from midterm to final exams, with Rasch analysis enabling the establishment of a consistent scale for tracking progress over time. Findings underscored the importance of aligning assessments with learning outcomes and ensuring fairness across diverse student groups. This research highlights the potential of Rasch analysis as a tool for improving the design, implementation, and evaluation of assessments in higher education, particularly in complex subjects like calculus. By integrating these insights, educators can refine assessment practices to better support student learning and achievement..

Keywords: Rasch analysis, educational assessment, fairness, student performance measurement, assessment validity

INTRODUCTION

Assessment plays a central role in higher education, particularly in courses that require a deep understanding of complex concepts, such as university-level calculus. For both instructors and students, fair and reliable assessment is crucial in ensuring that students' abilities are accurately measured and that educational outcomes are aligned with the learning objectives. Assessments not only measure current student performance but also shape their approach to lifelong learning, necessitating robust design strategies Boud and Falchikov (2006). However, one challenge that persists in many large-scale university courses is how to measure student progress in a way that is both equitable and consistent across diverse student subgroups, such as those studying engineering, economics and applied mathematics.

Traditional methods of assessment often struggle to address issues of fairness and reliability, particularly when it comes to the diversity of students' academic backgrounds and learning styles. Instructors face the complex task of designing exams that are capable of accurately measuring the range of skills students are expected to develop. The relationship between reliability and validity remains a cornerstone of educational

✉ shiralikadyrov@gmail.com **(*Correspondence)** ✉ Rosemary.Callingham@utas.edu.au ✉ a.makhmutova@newuu.uz ✉ g.pinto@newuu.uz

Contribution to the literature

- This article contributes to the current literature by demonstrating the application of Rasch analysis to evaluate the validity, reliability, and fairness of university-level calculus assessments, addressing gaps in ensuring equitable measurement across diverse student subgroups.
- It provides empirical evidence of student growth from midterm to final exams through a consistent measurement scale, enhancing the understanding of how assessments can track progress over time. By highlighting the alignment of assessment items with learning outcomes and identifying areas for pedagogical intervention, the study offers practical insights for improving assessment design in higher education.
- Additionally, it underscores Rasch analysis as a robust tool for ensuring fairness and precision in measuring complex skills, supporting its broader adoption in educational assessment research.

assessments, as they are interdependent in ensuring meaningful results (Moss, 1994). In such contexts, Rasch analysis offers a powerful tool for examining the validity, reliability, and fairness of assessments by providing a statistical framework that evaluates both the performance of individual students and the quality of exam items themselves.

The present study mainly aims to apply the Rasch analysis to the mid-term and final examinations of a calculus course at the university level. We also want to check the extent to which the questions in the exams appropriately measure the construct of student ability and how the consistency of this measurement holds across diverse student subgroups. We further investigate the reliability of these exams by providing stable measures of student performance across different groups. We will also discuss how linking items across the midterm and final exams enables the establishment of a consistent scale for monitoring student progress over time.

This study will address the following research questions (RQs):

RQ1. To what extent do the exam questions appropriately measure the intended construct of student ability, as determined through Rasch analysis, across diverse student subgroups?

RQ2. How reliable is the test in providing consistent measures of student ability across different student subgroups, as assessed using Rasch analysis?

RQ3. How can linking items between midterm and final exams help establish a consistent scale for measuring student progress over time?

By addressing these questions, this research aims to shed light on the potential of Rasch analysis as a tool for enhancing the fairness and reliability of assessments in higher education. Furthermore, it aims to offer actionable guidance for educators and policymakers wishing to improve assessment methodologies in large, diverse academic settings.

LITERATURE REVIEW

Rasch Analysis in Educational Assessment

Rasch analysis, named after Danish mathematician Georg Rasch, is a psychometric model used to evaluate and scale assessment data. In education, Rasch models help analyze test questions and student answers. They are one of the best ways to measure hidden traits, like how skilled a student is. Unlike simple scoring, Rasch analysis turns raw test scores into more detailed data. This makes it easier to measure both student ability and how hard a question is (Bond et al., 2020). The Rasch model is often used in big educational tests. It's helpful because it can handle differences in student skills and question difficulty in a clear and reliable way (Andrich, 1988). For example, Callingham and Bond (2006) explain how Rasch models are used in education to make large-scale testing more accurate.

In recent years, Rasch analysis has been in popular use on university-level examinations for those courses that involve the assessment of complex skills, such as calculus. It has been documented that Rasch models indicate possible problems of item difficulties and discriminations so that the exams measure what they are supposed to (Baartman et al., 2006). Besides, the Rasch analysis provides a better insight into the fairness of assessments by showing if some items disproportionately advantage the specific student subgroups, thus allowing instructors to modify it to be more equitable in testing (Birenbaum, 2007). For example, Meijer et al. (2020) investigate Rasch analysis in enhancing assessment literacy, an integral component in the facilitation of fair and valid assessment practices for diverse student populations. The study identifies the Rasch analysis process as crucial for uncovering the biases within the assessment toward enhanced fairness of the test.

In a related context, Callingham and Watson (2005) address the issue of measuring statistical literacy for which Rasch analysis is particularly effective. Callingham and Watson (2005) shows how Rasch models can yield a valid quantification of such complex student competencies and allow for a more detailed

understanding of student abilities, while also ensuring that assessments reflect the skills they are intended to measure. Along related lines, Vrikki et al. (2024) apply Rasch analysis to explore literacy achievement as an area where its relevance to educational assessment is underlined by the potential of the approach to monitor and assess students' progress in literacy, with significant implications for informing large-scale educational reforms.

Fairness and Validity in Assessments

Fairness in educational assessments is a major concern, especially in large university courses where students come from diverse backgrounds and have different learning experiences. Research has shown the importance of creating valid and fair assessments that give all students an equal chance to show their abilities, no matter their field of study or background (Liu & Boone, 2023). Validity means the assessment actually measures what it's supposed to measure, while fairness ensures every student has an equal opportunity to succeed (Messick, 1995). According to Gerritsen-van Leeuwenkamp et al. (2017), the quality of assessments in higher education depends on factors like validity, reliability, transparency, and fairness. These are essential to make sure assessments accurately and fairly measure student learning based on the intended goals.

In university calculus, fairness and validity are especially important because the subject often acts as a barrier for students entering fields like engineering or economics. Research highlights that assessments must consider the diversity of student strengths, learning styles, and problem-solving approaches (Boud & Falchikov, 2006). Traditional methods of analyzing test questions often fail to spot biases or differences in how questions work for different groups of students. This can make the assessment process unfair (Wilcox, 2011). A case study by Zheng et al. (2024) found a gap in the USA. higher education between what assessment practices are recommended and what actually happens. The study stresses the need for assessments that focus on students' varied needs and promote equal opportunities for learning and success. On the other hand, Rasch analysis serves to comprehensively appraise both the functionality of each single test item and the identification of any source that could cause bias in the said process Knight (2002); Linacre and Wright (2000). This method not only ensures that items function appropriately across different student populations but also helps identify item difficulty, aligning assessments with students' varying abilities.

Recent studies further underscore the importance of tailoring calculus assessments to address diverse student needs and enhance conceptual understanding. Illanes et al. (2025) analyzed derivative problems in engineering textbooks, identifying 21 subfields of problems and highlighting the need for targeted didactic strategies to support complex concept acquisition in engineering education. Similarly, Tatira (2025) used APOS theory to reveal undergraduate students' struggles with applying integral calculus to kinematics, particularly in higher-level integration and connecting concepts like displacement and velocity, emphasizing the need for assessments that bridge procedural and contextual understanding. Additionally, Aparicio-Landa et al. (2025) proposed an averages-based approach to teaching the fundamental theorem of calculus, finding it pedagogically effective in strengthening students' conceptual knowledge, which suggests potential for designing assessments that leverage such alternative approaches to improve learning outcomes.

Inclusion of Rasch analysis in the testing process will heighten the content and construct to be fair, and valid for the equity of students to show knowledge. Further, the results by Gerritsen-van Leeuwenkamp et al. (2017) and Zheng et al. (2024) offer emphasis on ensuring clarity and efficiency in the communication among the students, educators, and test specialists to clearly capture the multidimensional nature of assessment quality better. Moving forward, it is important that future research further investigates ways in which the application of the more advanced statistical methods the Rasch analysis aligns better with fairness in large-scale assessment of items of difficulty against student ability.

Use of Rasch Analysis for Linkage and Growth Measurement

One of the key strengths of Rasch analysis is its ability to connect assessments over time. This creates a clear scale for measuring student progress. This feature is especially useful in courses with multiple assessments, like midterms and finals, where instructors want to track how students improve throughout the semester. Rasch models can link questions from different tests, making it easier to measure student growth from one exam to the next (Wright, 1982). Rasch analysis does more than just compare scores. It provides a precise, data-driven way to measure student progress. For example, studies like Day et al. (2024) on multiplicative reasoning use Rasch analysis to track how students move through different stages of learning. This shows how growth can be measured on a continuous scale, similar to linking assessments over time.

Linking assessments is especially important in university calculus courses, where midterms and finals often have different formats and content. Using Rasch analysis to connect these assessments creates a continuous scale for measuring student progress. This is more accurate than just comparing raw scores. Research shows that linking questions this way can fix issues with varying question difficulty across exams and provide a more consistent way to measure student ability (Schuwirth & van der Vleuten, 2020; Van Der Vleuten & Schuwirth, 2005). It therefore connects with the wider

challenge, from within this set of selected works (Meijer et al., 2020), to secure construct validity within educational assessments that also arise with collaborative learning assessment literacy. Such work highlights an approach wherein an assessment aligns with what will be achieved through learning and agrees with Rasch models that set out the continued appropriate use of assessment items through time as learning objectives.

Moreover, Rasch models provide a powerful tool for monitoring growth at the individual student level, delivering detailed insights into how each student progresses over the course of their studies. This approach enables the provision of more tailored feedback, which is particularly valuable in large, diverse student cohorts where uniform measures of success often fall short (Linacre & Wright, 2000). More utility of Rasch analysis in learning progressions has been underlined by Day et al. (2024), underlining its ability to detect and remediate changes in student conceptions, comparable to nuanced ability tracking at a university calculus class. Furthermore, as pointed out by Meijer et al. (2020), growing demands for more effective assessment designs in higher education—especially under collaborative learning—underline the greater relevance of Rasch analysis. This methodology will not only identify areas for improvement in individual and group assessments but also ensure fairness and precision in the measurement of student development.

Item Types and Group Differences

The nature of the items in the assessment has a great deal to do with just how well the results of that assessment are able to reflect student ability across different populations. Items that measure concepts-appropriately assessing students' grasp of key principles- are different in keyways from items that require computation-appropriately measuring students' use of those principles. In university calculus, both types of items are in common use, and understanding their performance across different subgroups of students is important to ensure that assessment provides valid and fair outcomes. The work of Callingham and Watson (2012) points out how Rasch analysis can be applied to the measurement of different aspects of teacher knowledge in mathematics, including procedural understanding and pedagogical awareness. The same structure would apply to gauging the abilities of students in calculus-a strong avenue of determining the item difficulties and testing the fit between test items and student capability in ensuring both the conceptual and computation items are matched against diverse abilities in students (Beswick et al., 2012). Moreover, Johnson et al. (2024) highlight the role of quantitative graph reasoning in mathematical assessments, demonstrating that students' ability to interpret and select appropriate graphical representations significantly impacts their performance, which is particularly relevant in calculus where understanding function behavior and rate of change is essential. Similarly, Leavy et al. (2022) demonstrate how Rasch analysis can reveal variations in mathematics teaching efficacy beliefs among prospective teachers, particularly in their confidence with conceptual versus procedural tasks, emphasizing the importance of prior experiences in shaping efficacy—insights that parallel the challenges students face when engaging with different types of assessment items in calculus.

Research has shown that conceptual items may favor students who are more comfortable with abstract thinking, while computational items may benefit those who have strong procedural skills (Baartman et al., 2006). Differences in how student groups—like engineering, economics, and applied math students—perform on test questions can lead to uneven results. Some groups might do better on certain types of questions than others. Rasch analysis helps identify these differences, giving a clearer picture of how questions and student groups interact during an assessment (Bloxham et al., 2011; Bond et al., 2020). As Callingham (2015) points out, Rasch analysis can create detailed scales that compare question difficulty to student ability. This helps show how well questions work for different student groups. By uncovering patterns in student answers, Rasch analysis makes it easier to design fairer tests that meet students' needs.

Rasch analysis also investigates the degree to which each of the individual test questions is able to distinguish between students who possess different levels of skill. That is, good questions sharply separate students who know the material from those who don't. Poor questions fail to provide useful information about student performance (Bond et al., 2020). In studying the discrimination of questions, the Rasch analysis helps improve the design of tests. This ensures that the assessments are fair and valid for all students irrespective of their backgrounds. This notion is supported by Beswick et al. (2012), who note that questions should be set at an appropriate level of difficulty given the level of student performance. This makes the assessment system fair and transparent, which is particularly crucial when measuring understanding and skills in subjects like mathematics.

Contextual Studies in Mathematics Education

Numerous contextual studies have increasingly focused on the use of psychometric models, especially Rasch analysis, in evaluating the validity of the instruments in measuring students' competencies. These modern models respond to the pressing need for accurate, trustworthy, and unbiased evaluation systems that adequately cater to heterogeneous student groups. Research has shown that Rasch analysis in item response theory (IRT) can capture the multi-faceted nature of

items' functioning, and indeed capture how different items are performing across different subgroups of students in areas such as engineering, economics, and applied mathematics (Di Nisio, 2010; Saidfudin et al., 2010; Taylor et al., 2020; Wei et al., 2014).

In Rasch analysis in engineering education, Saidfudin et al. (2010) considered the classification of students for reporting purposes with respect to defined outcomes and the construction of logit rulers as a measurement of students' abilities as great innovations. From this, it was concluded that Rasch can assist researchers in determining areas of construct underrepresentation that particularly aid in the construction of valid and reliable instruments for assessment. Di Nisio (2010) also demonstrated how Rasch analysis is applied in mathematics education by highlighting how it transforms raw scores to interval-level measurements, thereby providing a precise profiling of student performance coupled with the objectivity and reliability of results.

Wei et al. (2014), in their analysis of the mathematics interest inventory, assessed the impact of IRT and cultural and social desirability aspects of the given context. This highlights the need to be attentive to various contextual phenomena when constructing tests to achieve equity and accuracy. Similarly, Lei et al. (2022) studied the cultural embedding of mathematics test items in China's national college entrance examination and how culture constructs students' response patterns and the character of mathematical issues.

Rasch methodology, through Boone's (2016) lens, is a powerful tool for model application in the refinement of assessment instruments, identification and remedy of invalid items, and general validity improvement. This is in close support to Kieftenbeld et al.'s (2011) studies which were able, through their psychometric tests, to assist in the formulation of practical educational measures. In the same fashion, Taylor et al. (2020) demonstrate the iterative design and development process of the tools with targeted outcomes of learning achievement using Rasch analysis on the bio calculus assessment. Elizar and Khairunnisak (2020) accentuate the additional strength of the Rasch model which allows for detailed analysis of mathematics assessment.

METHODOLOGY

Participants

The dataset comprises 369 students enrolled in a university-level calculus course, divided across multiple academic disciplines. The students belong to various programs such as freshman applied mathematics (FAM), freshman artificial intelligence and robotics (FAR), freshman engineering (FE), freshman cyber security (FCS), freshman economics and data science (FED), freshman pedagogy (FP), and freshman software engineering (FSE). **Figure 1** provides a detailed breakdown of the number of students in each subgroup.

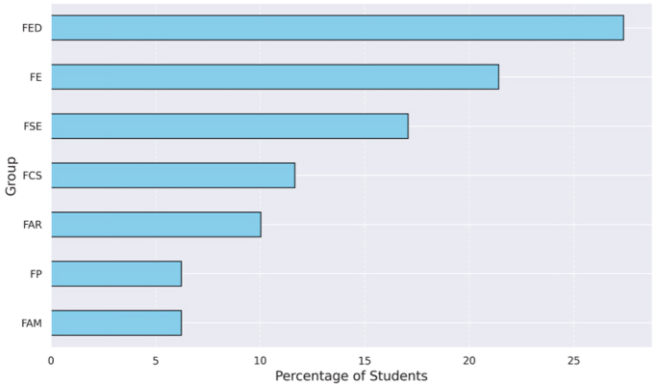

Figure 1. Percentage distribution of students across groups (Source: Authors' own elaboration)

Data Collection

Assessments

The assessments in this course were collaboratively prepared by four professors with extensive teaching expertise in calculus, ensuring a comprehensive evaluation of students' understanding of the subject. Both the midterm, conducted in week 7, and the final exam, held in week 16, were designed to align with the syllabus learning outcomes and assess a balance of conceptual understanding, computational skills, and the ability to apply calculus principles. The midterm, a 90-minute exam comprising 17 items, focused on foundational topics such as limits, continuity, and differentiation, with tasks ranging from evaluating limits using the epsilon-delta definition to providing the continuity of piecewise functions and solving optimization problems through differentiation. The final exam, a 120-minute assessment with 19 items, expanded on these topics to include integration and its applications, such as calculating definite and improper integrals, finding areas and volumes of revolution, and interpreting the average value of functions. Advanced problems required students to analyze critical points, concavity, and asymptotes and synthesize these findings in curve sketching. To provide a fair and thorough evaluation of student performance, both assessments were scored using a partial credit approach. This method acknowledged students' partial understanding and rewarded their progress on intermediate steps, ensuring that their problem-solving efforts were accurately reflected in their scores. **Figure 2** illustrates the 15-week course timeline, showing the strategic placement of the midterm and final exams to assess learning progress.

Item mapping

Figure 3 illustrates the alignment of midterm and final exam items, where nodes represent individual items, labeled with a prefix "M" for midterm (e.g., MQ1a

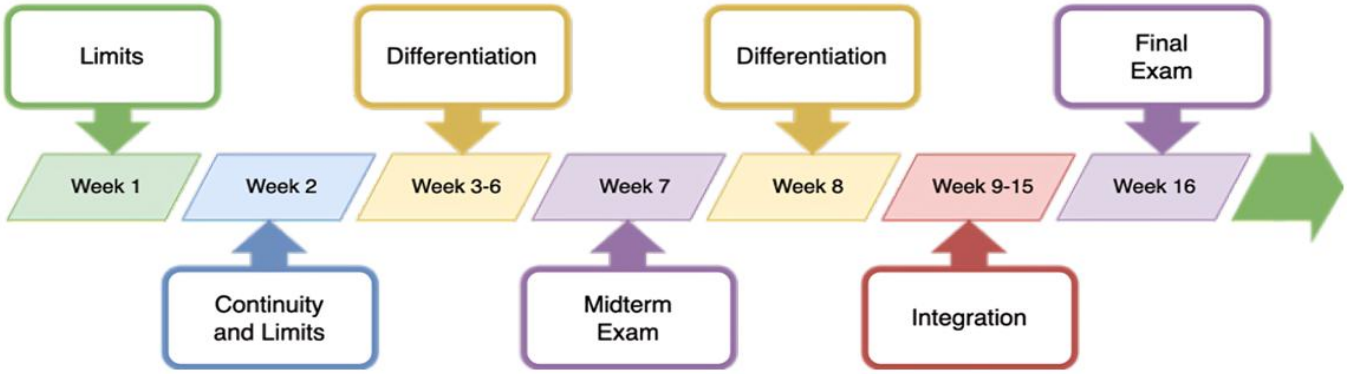

Figure 2. Timeline of course structure and assessment placement (Source: Authors' own elaboration)

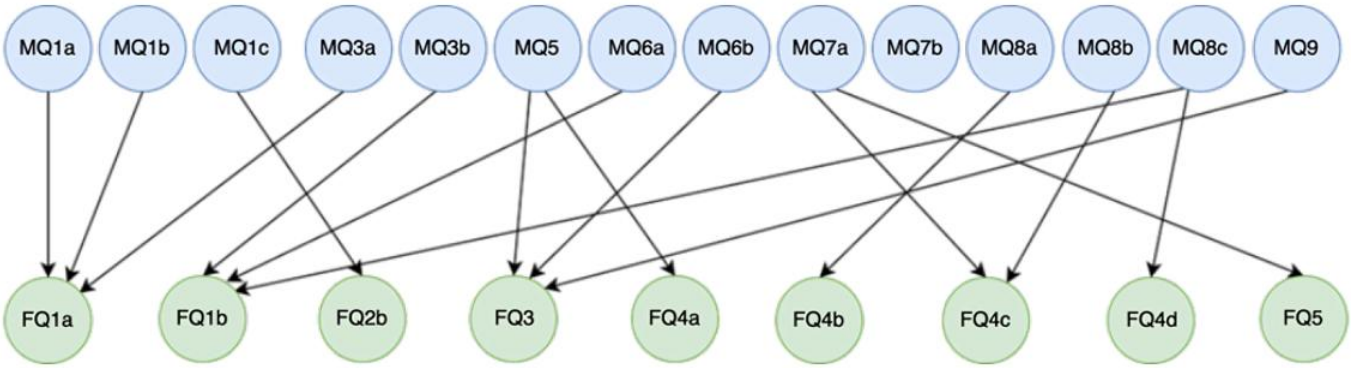

Figure 3. Conceptual alignment of midterm exam items to final exam items (Source: Authors' own elaboration)

stands for midterm question 1, part a) or "F" for final (e.g., FQ2b for final question 2, part b). The parts within an item either represent sequential steps in problem-solving or group subproblems under a common topic within the same exam. Edges between nodes indicate conceptual alignments identified through careful manual inspection by the authors. Items were aligned if they addressed analogous topics and required similar solution techniques and knowledge to solve. Unlinked items, which did not meet these criteria, were excluded from **Figure 3**. This alignment emphasizes the progression and continuity of key calculus concepts across the two assessments, facilitating a structured evaluation of student learning.

Data collected

Data for the study were collected through student exam scores. For each participant, raw scores for both the midterm and final exams were recorded. The primary data for this study were collected from two assessments: a midterm exam and a final exam. Both assessments were designed to evaluate students' understanding of key calculus concepts. The midterm focused on limits, continuity, and differential calculus, while the final exam included these topics along with integration.

The midterm exam consisted of 25 questions, split into two sections: 15 conceptual questions that tested students' understanding of fundamental principles, and 10 computational questions that assessed their ability to apply these principles in problem-solving scenarios. The final exam, in contrast, included 30 questions, 18 of which were conceptual and 12 computational. Both exams were designed to assess a range of difficulty, from basic recall to more complex application-based problems. The items were scored using a partial credit approach.

Each item in the exams was constructed to align with the course's intended learning outcomes. Additionally, both exams were designed to measure the same underlying construct of "student ability" in calculus, ensuring consistency between the two assessments.

Rasch Analysis Procedure

Rasch analysis places all items and test-takers (persons) on a single scale. This apapproach is underpinned by three assumptions:

1. All items work together to measure a single, unidimensional construct. This construct may be "thick", such as mathematics, or narrower, such as calculus, but every item should contribute to the defined form.
2. All items are independent of each other. This independence means that a correct response to one item may not be a prerequisite for a correct response to another item.

Table 1. Interpretation of mean-square fit statistics

Fit value	Implication for measurement
> 2.0	Distorts or degrades the measurement system
1.5-2.0	Unproductive for construction of measurement, but not degrading 0.5
1.5	Productive for measurement
< 0.5	Less productive for measurement, but not degrading

3. A higher score implies more of the construct, that is the construct is hierarchical and additive, developing from low to high capacity (Smith, 2001).

It is worth noting that these assumptions underpin all assessments, but they are made explicit in a Rasch analysis. Smith (2001) argued that if these assumptions are met, the instrument provides a valid assessment. These assumptions can be tested through the fit of the data to the Rasch model. Unlike other forms of model building, in Rasch analysis the model is paramount, and the data are fitted to the model rather than the model being modified to fit the data. The notion of fitness will be discussed further below.

In this study, the analyses were conducted in two phases. First, the measurement scale was established, second, the scale was "anchored" and person measures were obtained for the midterm and final assessments using this anchored scale to ensure that they were directly comparable.

To establish the measurement scale, the initial step was to conduct a straightforward analysis of the midterm and final tests using Winsteps Rasch Measurement 5.2.0.0 (Linacre, 2022) using the Rasch partial credit model. This analysis established the extent to which the assumptions were met for each test. Fit to the Rasch model was evaluated at both the overall test level and the item level using infit and outfit mean-square (MNSQ) statistics, together with their standardized z (ZSTD) values. Infit is an information-weighted fit statistic that is most sensitive to unexpected responses on items targeted near a student's estimated ability, whereas outfit is outlier-sensitive (unweighted) and is more influenced by unexpected responses on very easy or very difficult items. For MNSQ, the expected value is 1.00, and for ZSTD the expected value is 0.00; values above 1.00 indicate more variation than the model expects (underfit), while values below 1.00 indicate overly predictable responses (overfit). **Table 1** provides a practical guide for interpreting infit and outfit values. Because Rasch estimation is probabilistic, exact fit is rarely observed; therefore, the ranges in **Table 1** are used as rule-of-thumb criteria for diagnosis. In particular, MNSQ values > 2.0 may distort measurement and threaten construct validity (Linacre, 2022).

Reliability statistics are also provided by Rasch analysis. These take the form of item and person separation reliabilities. Items and persons are distributed across the measure and the separation reliabilities indicate how stable or reproducible these locations are. In general, person reliability should be > 0.8 and the separation should be > 2.0. Low item reliabilities (< 0.5) indicate that the sample size is too small to obtain an accurate measure (Linacre, 2022). The same statistics apply to both person measures and item measures. Taken together, these statistics provide an indication of the validity of the assessment to measure the intended construct (Bond et al., 2020). The next step was to "equate" the two assessments. There are two forms of test equating—a common item or a common person (Griffin & Callingham, 2006). In this study, there were no common items, but the same 369 students undertook each assessment. The combined data file, with responses of each student to every item, was Rasch analyzed using Winsteps Rasch Measurement 5.2.0.0 (Linacre, 2022), following exactly the same process as for the two previous analyses.

To enable direct comparison of students' performances across time the midterm and final assessments were analyzed anchored to the difficulty level of the items determined by the merged analysis, using the procedure detailed in Winsteps (Linacre, 2022). In this way, performances by the students are measured against the same scale, regardless of when the test was taken, and could be directly compared Bond et al. (2020). Person measures (termed abilities in Rasch analysis) in logits were obtained for every student at the two time points. Several analyses were then undertaken.

FINDINGS

Overall summary statistics for the midterm and final assessments are shown in **Table 2**.

In terms of model fit, the person and item mean infit/outfit values for the midterm, final, and combined analyses were close to 1.00 (**Table 2** and **Table 3**), indicating that the overall response patterns closely matched Rasch model expectations. Importantly, none of the item-level fit statistics exceeded thresholds

Table 2. Summary statistics for midterm and final assessments

Assessment	n	Infit	z infit	Outfit	z outfit	Separation	Reliability	Cronbach's alpha
Midterm persons	369	1.05	0.10	1.07	0.12	2.89	0.89	0.90
Midterm items	17	1.02	0.01	1.07	-0.04	10.39	0.99	-
Final persons	369	1.11	0.18	1.11	0.16	3.21	0.91	0.90
Final items	19	1.05	0.05	1.11	0.11	13.18	0.99	-

Note. "-" indicates that Cronbach's alpha was not calculated for item-level statistics

Table 3. Summary statistics for the combined measure

Assessment	n	Infit	z infit	Outfit	z outfit	Separation	Reliability	Cronbach's alpha
Combined persons	369	1.08	0.20	1.12	0.21	4.12	0.94	0.94
Combined items	36	1.03	0.03	1.13	0.25	11.82	0.99	-

Note. "-" indicates that Cronbach's alpha was not calculated for item-level statistics

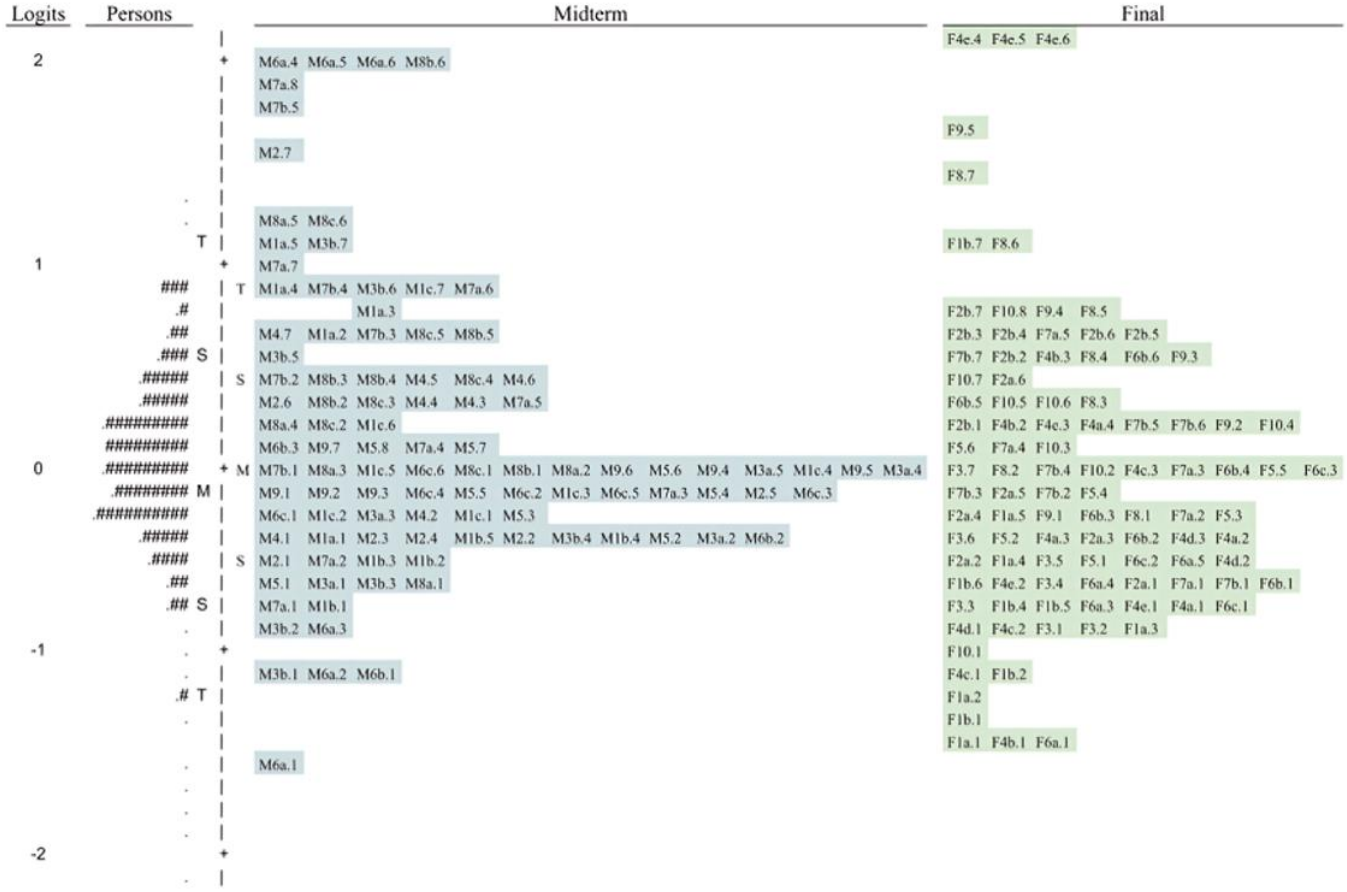

Figure 4. Wright map of all items analyzed together (Source: Authors' own elaboration)

typically associated with construct-threatening underfit (e.g., MNSQ > 2.0). This suggests that responses were not dominated by excessive randomness (guessing, ambiguous wording, or multidimensionality) and supports the validity of interpreting the resulting person measures as a coherent estimate of calculus ability on a single scale.

The assessments were also examined at item level but in neither test did any item misfit sufficiently to threaten measurement. Hence all items were included in the analyses. The two assessments provided valid and reliable measures of the underlying construct.

Table 3 presents the summary statistics for the combined measure, including fit statistics (infit and outfit), separation, reliability, and Cronbach's alpha for both persons and items. These statistics indicate the validity and reliability of the combined assessment, ensuring that the scale provides consistent and meaningful measures across both the midterm and final assessments.

To interpret the construct, the analysis provides a measure of the difficulty of each item in logits, the unit of Rasch measurement. The distribution of the items can be shown graphically as a Wright map. **Figure 4** shows the Wright map for the combined analysis. For comparison, the midterm (LHS) and final (RHS) items have been separated.

The items are labeled M (midterm) or F (final) followed by the item number in the test, with the score code following the point. Hence, F4b.2 indicated the final test, Item number 4b, score-code 2. The most difficult items are at the top of the scale and the easiest at the bottom. The person distribution is also shown with the # indicating 3 persons and a . indicating 1 or 2 persons.

There are several aspects that are worth noting. The items are spread out across the scale in both assessments, indicating that the tests have a similar range of difficulty. The mean item difficulty is constrained by the analysis to 0.00 and, in this map, the mean person measure is very close to that of the items.

Table 4. Overall mean difference between midterm and final assessments (n = 369)

	Mean ability	Standard deviation	t	p
Midterm	-0.157	0.62	3.34	0.001
Final	-0.085	0.73		

Taken together, these observations suggest good quality instruments that provide sufficient opportunity for all students from the least to the most able to demonstrate what they know and understand. Further, the placement of each person indicates that they have a 50% probability of answering correctly on items at the same position on the scale. That is, students whose ability matches the item difficulty level (0.00 logit) have a 50% probability of answering correctly on any items in the row, such as M7b.1, F3.7, and so on Siemon and Callingham (2019).

Using the common scale established through the anchored item difficulties, changes in student performance across the midterm and final assessments were evaluated. Mean overall performance on the midterm and final tests showed a statistically significant difference on a paired sample t-test. Details are shown in **Table 4**.

The students were clustered into different courses and means and standard deviations were calculated for each grouping at the two time points, with the results summarized in **Table 5**.

Table 5. Summary statistics for course groups

Course	n	Midterm M	Midterm SD	Final M	Final SD
FAM	23	-0.03	0.90	-0.02	1.01
FAR	37	-0.04	0.39	0.09	0.31
FCS	43	-0.07	0.40	-0.05	0.53
FE	79	-0.21	0.64	-0.11	0.81
FED	101	-0.23	0.66	-0.08	0.57
FP	23	-0.35	0.67	-0.37	0.80
FSE	63	-0.07	0.62	-0.12	0.95

Note. M: Mean & SD: Standard deviation

These findings indicate that the groups performed differently, with some showing growth and others apparently declining in performance. **Figure 5** shows the mean performances of each group at different time points.

Clearly, the FP group were the least able students, and their performance did not change over time. In contrast, the FAR group was not only highly able students but also showed considerable growth. The FE group appeared to have declined in performance over time. To check whether the tests had an inbuilt bias to these groups, a one-way ANOVA was performed using the anchored midterm and final person abilities as the dependent variables. Despite the differences in performance, there were no statistically significant differences among any group on either measure (midterm df = 6, F = 1.347, p = .235; final df = 6, F = 1.014,

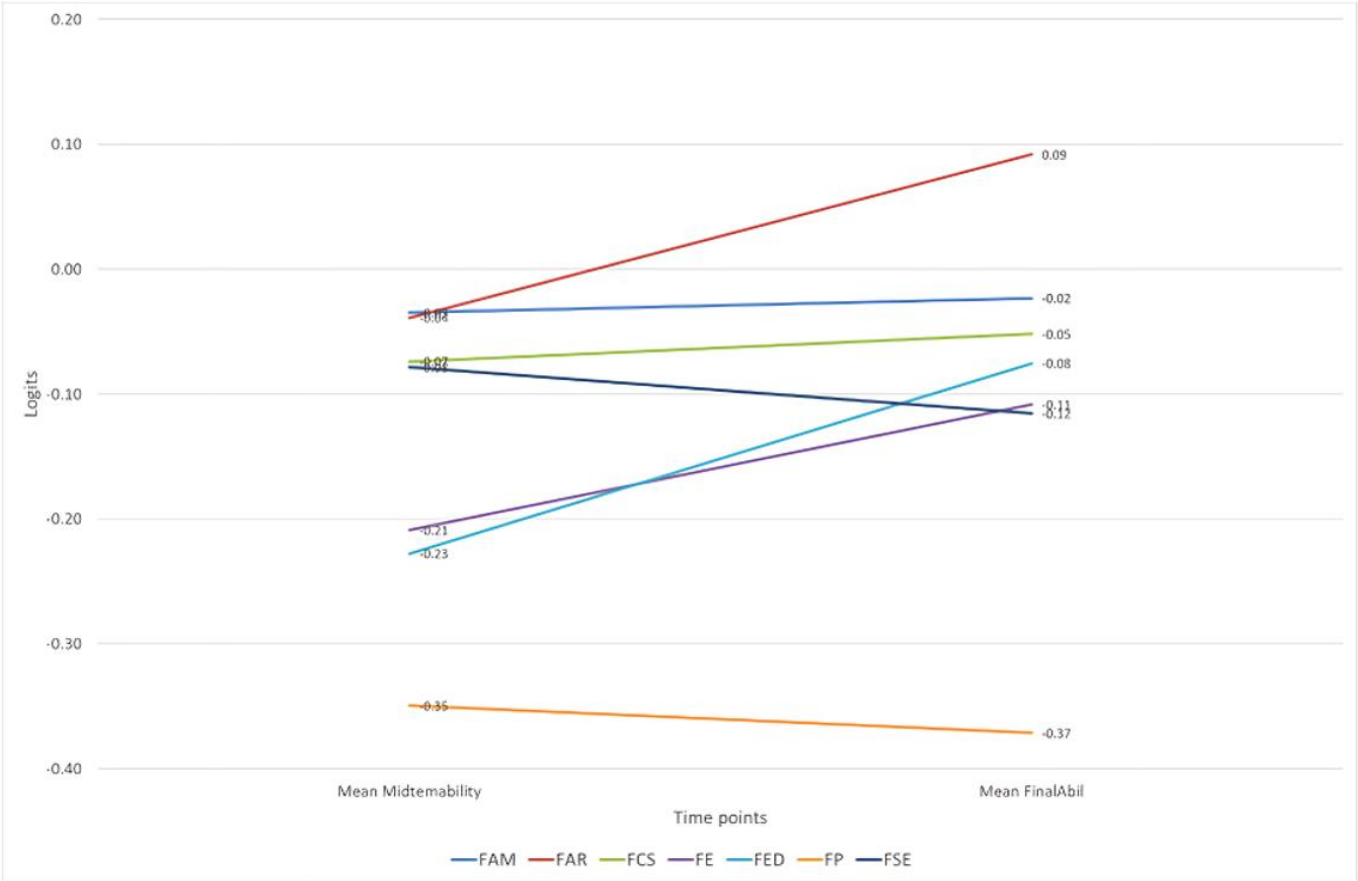

Figure 5. Performance on midterm and final assessments by group (Source: Authors' own elaboration)

p = .416). These findings indicate that the tests are working in the same way across all groups and are measuring real differences among the groups.

DISCUSSION

These results have important implications for the de

sign, implementation, and evaluation of university-level calculus assessments in terms of how validly they can be said to measure student learning outcomes. Using Rasch analysis allows us to test the validity and reliability of assessments rigorously, and thus this is an evident and evidence-framed framing of how well these tests represent student ability.

First of all, our analyses addressed the key question of how well the exam questions measure the intended construct of student ability. The descriptive statistics indicate that both sets of midterm and final exams were valid and reliable measures of this construct. Finding both high separation and reliability indices for items and persons, the tests were effectively distinguishing between higher and lower levels of student performance. The close approximation of the mean item difficulty to that of the student's ability serves to further confirm that the tests were well-set, catering to students across the spectrum of ability. This addresses the intention of designing assessments that measure what they intend to do since it shall be appropriate for a wide range of students.

Next, the reliability of the assessments across various student subgroups was carefully considered. The Rasch analysis indicated that both the midterm and final exams provided consistent measures of student ability, regardless of subgroup. The reliability statistics for both persons and items were strong, and the absence of significant differences between groups in terms of their ability measures suggests that the tests performed equally well across diverse student backgrounds. This highlights the consistency and fairness of the assessments, addressing concerns about how well they function across different groups.

Lastly, our **RQ3** focused on how linking items between the midterm and final exams could establish a consistent scale for measuring student progress over time. The findings show that the use of anchored items allowed us to create a common scale, which helped track student progress between the two assessments. The significant difference in student performance between the midterm and final exams, as shown in the paired sample t-test, provides evidence of real growth over time. The Wright map also showed that the items from both tests were appropriately placed to make a unidimensional scale, thereby demonstrating coherence in the scale and, hence, students' developing competence. This continuity in the design of the assessment is critical for determining how students develop their understanding during the course of a semester.

The results thus imply that at a more detailed instructional level, areas such as advanced techniques of integration and curve sketching were difficult for students. These results ensure that targeted interventions, such as practice sessions or focused tutorials, are implemented to help students overcome these challenges. Assessments that are closely aligned with the learning outcomes also ensure that better-informed teaching strategies are directed at fostering meaningful student progress.

This study has provided strong evidence for the validity and reliability of the assessments, but there are some limitations. First, this dataset is drawn from a single institution, which may reduce the generalizability of results. In fact, this approach might be expanded to other institutions or disciplines in further research in order to determine if these findings would hold across those other settings. Furthermore, while assessment alignment and student performance were accounted for, further research into the effectiveness of specific teaching strategies could provide a broader, deeper perspective on how instruction truly impacts students.

CONCLUSION

This study actually points out the important role of Rasch analysis in maintaining the validity and reliability of educational assessments. In our analyses, we have used university calculus exams to show how highly aligned and thoughtfully constructed assessments can measure varied student abilities while supporting progress in conceptual learning. The findings point out the importance of aligning assessment to learning outcomes and allow the instructor to locate any break in understanding which then can be treated with specific remediation.

Although the results are strong, reliance on one dataset from a single institution limits generalizability. Future research should extend this methodology across varied contexts and disciplines to validate its broader applicability. Further, the interaction between teaching strategies and student performance could be investigated for an even more comprehensive view of how assessments contribute to learning. Another promising direction for future work is in understanding how different item types based on specific calculus topics work together to support inferences of student abilities across a range of disciplines. Rasch analysis could be used to further evaluate these item types for insight into their role in the measure and its support of diverse learning trajectories. By integrating these insights, educational institutions can refine their practices to improve both teaching and learning outcomes in complex domains such as calculus.

Author contributions: SK: project administration, investigation, writing – review & editing; **GP:** investigation, data curation, writing – review & editing; **AM:** writing – original draft, writing – review & editing, funding acquisition; **RC:** conceptualization, methodology, formal analysis, supervision, validation. All authors agreed with the results and conclusions.

Funding: No funding source is reported for this study.

Acknowledgements: The third author acknowledges the funding by the Ministry of Science and Higher Education of the Republic of Kazakhstan under Project No. AP25796179.

Ethical statement: The authors stated that the study was approved by the Research and Ethics Review Committee of New Uzbekistan University (Department of Science, Tashkent, Uzbekistan) (Approval No. 41, dated April 15, 2025). The research involved secondary analysis of anonymized student assessment data collected as part of routine educational practice. No experimental intervention was conducted, and no personally identifiable information was used. The requirement for written informed consent was formally waived by the Committee due to the retrospective and fully anonymized nature of the dataset. All data were handled in accordance with institutional standards for confidentiality and privacy.

AI statement: The authors stated that they used ChatGPT (OpenAI, GPT-4o0 for limited language editing and proofreading. All AI-assisted content was reviewed and approved by the authors, who remain fully responsible for the manuscript.

Declaration of interest: No conflict of interest is declared by the authors.

Data sharing statement: Data supporting the findings and conclusions are available upon request from the corresponding author.

REFERENCES

Andrich, D. (1988). *Rasch models for measurement*. SAGE. https://doi.org/10.4135/9781412985598

Aparicio-Landa, E., Sosa-Moguel, L., García-Almeida, G., & Avila-Vales, E. (2025). University students and professors' reflections on an averages-based approach to the fundamental theorem of calculus. *Eurasia Journal of Mathematics, Science and Technology Education, 21*(3), Article em2593. https://doi.org/10.29333/ejmste/16002

Baartman, L. K., Bastiaens, T. J., Kirschner, P. A., & Van der Vleuten, C. P. (2006). The wheel of competency assessment: Presenting quality criteria for competency assessment programs. *Studies in Educational Evaluation, 32*(2), 153-170. https://doi.org/10.1016/j.stueduc.2006.04.006

Beswick, K., Callingham, R., & Watson, J. (2012). The nature and development of middle school mathematics teachers' knowledge. *Journal of Mathematics Teacher Education, 15*, 131-157.

Birenbaum, M. (2007). Evaluating the assessment: Sources of evidence for quality assurance. *Studies in Educational Evaluation, 33*(1), 29-49. https://doi.org/10.1016/j.stueduc.2007.01.004

Bloxham, S., Boyd, P., & Orr, S. (2011). Mark my words: the role of assessment criteria in UK higher education grading practices. *Studies in Higher Education, 36*(6), 655-670. https://doi.org/10.1080/03075071003777716

Bond, T., Yan, Z., & Heene, M. (2020). *Applying the Rasch model: Fundamental measurement in the human sciences*. Routledge. https://doi.org/10.4324/9781315814698

Boone, W. J. (2016). Rasch analysis for instrument development: Why, when, and how? *CBE – Life Sciences Education, 15*(4), Article rm4. https://doi.org/10.1187/cbe.16-04-0148

Boud, D., & Falchikov, N. (2006). Aligning assessment with long-term learning. *Assessment & Evaluation in Higher Education, 31*(4), 399-413. https://doi.org/10.1080/02602930600679050

Callingham, R. (2015). Measurement challenges in mathematics education research. In *Handbook of international research in mathematics education* (pp. 462-480). Routledge.

Callingham, R. C., & Bond, T. G. (2006). Research in mathematics education and Rasch measurement. *Mathematics Education Research Journal, 18*(2), 1–10. https://doi.org/10.1007/BF03217432

Callingham, R., & Watson, J. M. (2005). Measuring statistical literacy. *Journal of Applied Measurement, 6*(1), 19-47.

Day, L., Siemon, D., Callingham, R., & Seah, R. (2024). Connecting the threads: The role of multiplicative thinking in algebraic, geometrical, and statistical reasoning. *Research in Mathematics Education, 26*(2), 325-347.

Di Nisio, R. (2010). Measure school learning through Rasch analysis: The interpretation of results. *Procedia-Social and Behavioral Sciences, 9*, 373-377. https://doi.org/10.1016/j.sbspro.2010.12.167

Elizar, E., & Khairunnisak, C. (2020). An introduction to the Rasch Measurement Model: A case of mathematics education students comprehensive test. *Kalamatika: Jurnal Pendidikan Matematika, 5*(1), 51-60. https://doi.org/10.22236/KALAMATIKA.vol5no1.2020pp51-60

Gerritsen-van Leeuwenkamp, K. J., Joosten-ten Brinke, D., & Kester, L. (2017). Assessment quality in tertiary education: An integrative literature review. *Studies in Educational Evaluation, 55*, 94-116. https://doi.org/10.1016/j.stueduc.2017.08.001

Griffin, P., & Callingham, R. (2006). A 20-year study of mathematics achievement. *Journal for Research in Mathematics Education, 37*(3), 167-186.

Illanes, M. K. G., Breda, A., Alvarado-Martínez, H., & Sala-Sebastià, G. (2025). Characterization of sub-fields of derivative problems in engineering textbooks. *Eurasia Journal of Mathematics, Science and Technology Education, 21*(3), Article em2591. https://doi.org/10.29333/ejmste/15987

Johnson, H. L., Donovan, C., Knurek, R., Whitmore, K. A., & Bechtold, L. (2024). Proposing and testing a model relating students' graph selection and graph reasoning for dynamic situations. *Educational*

Studies in Mathematics, 115(3), 387-406. https://doi.org/10.1007/s10649-024-10299-4

Kieftenbeld, V., Natesan, P., & Eddy, C. (2011). An item response theory analysis of the mathematics teaching efficacy beliefs instrument. *Journal of Psychoeducational Assessment, 29*(5), 443-454. https://doi.org/10.1177/0734282910391062

Knight, P. T. (2002). Summative assessment in higher education: Practices in disarray. *Studies in Higher Education, 27*(3), 275-286. https://doi.org/10.1080/03075070220000662

Leavy, A., Bjerke, A. H., & Hourigan, M. (2023). Prospective primary teachers' efficacy to teach mathematics: Measuring efficacy beliefs and identifying the factors that influence them. *Educational Studies in Mathematics, 112*(3), 437-460. https://doi.org/10.1007/s10649-022-10181-1

Lei, P., Kong, W., Han, S., Lv, S., & Wang, X. (2022). The mathematical culture in test items of national college entrance examination in China from 1978 to 2021. *Mathematics, 10*(21), Article 3987. https://doi.org/10.3390/math10213987

Linacre, J. M. (2022). Winsteps Rasch measurement v. 5.2.0.0. *Winsteps*. https://www.winsteps.com

Linacre, J. M., & Wright, B. D. (2000). Winsteps. *Winsteps*. http://www.winsteps.com/index.htm

Liu, X., & Boone, W. J. (2023). *Advances in applications of Rasch measurement in science education*. Springer. https://doi.org/10.1007/978-3-031-28776-3

Meijer, H., Hoekstra, R., Brouwer, J., & Strijbos, J.-W. (2020). Unfolding collaborative learning assessment literacy: A reflection on current assessment methods in higher education. *Assessment & Evaluation in Higher Education, 45*(8), 1222-1240. https://doi.org/10.1080/02602938.2020.1729696

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749. https://doi.org/10.1037/0003-066X.50.9.741

Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher, 23*(2), 5-12. https://doi.org/10.3102/0013189X023002005

Saidfudin, M., Azrilah, A., Rodzo'An, N., Omar, M., Zaharim, A., & Basri, H. (2010). Use of Rasch analysis to measure students performance in engineering education. In *Proceedings of the 7th WSEAS International Conference on engineering education* (pp. 435-441). WSEAS.

Schuwirth, L. W., & van der Vleuten, C. P. (2020). A history of assessment in medical education. *Advances in Health Sciences Education, 25*(5), 1045-1056. https://doi.org/10.1007/s10459-020-10003-0

Siemon, D., & Callingham, R. (2019). Researching mathematical reasoning: Building evidence-based resources to support targeted teaching in the middle years. In *Researching and using progressions (trajectories) in mathematics education* (pp. 101-125). Brill.

Smith Jr, E. V. (2001). Evidence for the reliability of measures and validity of measure interpretation: A Rasch measurement perspective. *Journal of Applied Measurement, 2*(3), 281-311.

Tatira, B. (2025). Undergraduate students' understanding of the application of integral calculus in kinematics. *Eurasia Journal of Mathematics, Science and Technology Education, 21*(3), Article em2601. https://doi.org/10.29333/ejmste/16049

Taylor, R. T., Bishop, P. R., Lenhart, S., Gross, L. J., & Sturner, K. (2020). Development of the biocalculus assessment (BCA). *CBE – Life Sciences Education, 19*(1), Article ar6. https://doi.org/10.1187/cbe.18-10-0216

Van Der Vleuten, C. P., & Schuwirth, L. W. (2005). Assessing professional competence: From methods to programmes. *Medical Education, 39*(3), 309-317. https://doi.org/10.1111/j.1365-2929.2005.02094.x

Vrikki, M., Kyriakides, L., & Dimosthenous, A. (2024). The potential of following-up inter-national large-scale assessment studies: Using PISA 2018 to develop a comprehensive model of effective teaching. *Educational Research and Evaluation, 30*(3), 249-273. https://doi.org/10.1080/13803611.2024.2344094

Wei, T., Chesnut, S. R., Barnard-Brak, L., Stevens, T., & Oliv´arez Jr, A. (2014). Evaluating the mathematics interest inventory using item response theory: Differential item functioning across gender and ethnicities. *Journal of Psychoeducational Assessment, 32*(8), 747-761. https://doi.org/10.1177/0734282914540449

Wilcox, R. R. (2011). *Introduction to robust estimation and hypothesis testing*. Academic Press. https://doi.org/10.1016/B978-0-12-386983-8.00001-9

Wright, B. D. (1982). *Rating scale analysis*. Measurement, Evaluation, Statistics, and Assessment Press.

Zheng, Y., Van Vliet, F., & Jin, J. I. (2024). Case study of the use of learner-centered assessment in the math school of a large university in the United States. *Educational Research and Evaluation, 30*(4), 476-494. https://doi.org/10.35542/osf.io/7gzrk