DOI: https://doi.org/10.64428/rm/v6.i2.5
Submission Date: 2026-04-20
Acceptance Date: 2026-06-16
Publication Date: 2026-06-30

Volume 6, Issue 2, June 2026, PP. 60-70
ISSN: (O) 2959-1821 (P) 2959-1813
Website: https://researchmosaic.com
Email: rm@researchmosaic.com

Psychometric Evaluation of a Fifth-Grade Standardized Examination: An Item Analysis of Multiple-Choice Questions

Tahir Ullah[*]**,** *Department of Education, Abasyn University, Peshawar, Pakistan*
Farzana Salim, *Department of Education, Abasyn University, Peshawar, Pakistan*
Jawad Ali Shah, *UET Peshawar, Pakistan*

Keywords

Classical Test Theory, Difficulty Index, Discriminating Power, Item Analysis, Standardized Examination, Primary Education, Psychometrics.

Abstract

The present research study has conducted an all-inclusive item analysis of the fifth-class standardized examination to assess its psychometric quality. To analyze the difficulty index (P) and the discriminating power (D) for fifteen MCQs, established on the answers from the higher-achievers and lower-achievers groups, the classical test theory was employed. It has been found that most of the items, i.e., 12 out of 15 MCQs, fell into the moderately difficult range (0.30<P<0.70), and three items were easy, having a P value higher than 0.70. It was also found that out of 15 items, ten items were in the "very good" discriminating category (D ≥ 0.40), while the rest of five items were either borderline or acceptable (0.20 ≤ D < 0.30) and needed valuation according to the discrimination investigation. The evaluation as a whole has an excellent reliability index value of 0.989 for distinguishing the high-scoring students from low-scoring students. This work highlights the significance of post-administration item analysis in confirming and refining classroom assessment tools.

INTRODUCTION

Evaluation in education is very crucial for assessing student learning, apprising teaching approaches, and making informed decisions regarding educational progress (Kumar et al., 2021). Standardized tests, particularly those that employ multiple-choice questions (MCQs), are still extensively utilized in this discipline due to their effectiveness, scalability, and impartial scoring (Rock et al., 2026). However, the validity and quality of a test cannot be guaranteed by the act of administering it. The utility of an evaluation is shown by its psychometric topographies, which are the arithmetical features that describe how efficiently it measures the proposed notion (Bandalos, 2018). For evaluations in primary-level education, such as fifth-class examinations, it is important to guarantee appropriate psychometric quality. At critical stages in a learner's academic route, these evaluations are frequently employed to direct involvement, identify learning gaps, and contribute to cumulative appraisals (COnsolata, 2024). A not-well-built test tool may lead to improper decisions about learners' ability and skill that could result in prejudiced educational judgments, as asserted by Bennett (2015).

Item analysis, a set of quantifiable approaches grounded on Classical Test Theory (CTT), provides vital tools for this assessment (Black & Wiliam, 2018). Two of the most central measures are the "Difficulty Index (P)" and the "Discriminating Power (D)." The difficulty index specifies how simple or complex a test tool is for the proposed viewers. It is considered the fraction of test-takers who deliver a correct answer. The optimum mark of difficulty is

[*]Corresponding author e-mail address: mostas85@gmail.com

described by the objective(s) of the test; for norm-referenced analyses predestined to allocate pupils, a *P*-value of unevenly 0.50 is frequently considered a good one; however, a range of 0.30-0.70 is usually acceptable (Brookhart & Nitko, 2019). The discriminating power of an item describes its ability to effectively differentiate between high-performing and low-performing learners. The discriminating power of an item is calculated by comparing the ratio of right responses in a high-scoring group (upper group) with a low-scoring group (lower group). High positive discrimination authenticates an item's alignment with the test objective, demonstrating that learners who grasp the material are more likely to answer correctly, while those who do not tend to answer incorrectly (Rock et al., 2026).

Item analysis procedures are not consistently applied in the development and evaluation of school-level and classroom-based standardized tests (Brown, 2019). Hypothetically defective tools are frequently employed because many instructors lack the resources or skill required to carry out such studies. The current study objects to diminishing this gap by employing item analysis on a particular fifth-class standardized test. By cautiously calculating and interpreting the difficulty index (*P*-values) and discriminating power (D-value) for every item included in the test, this study sought to assess the overall psychometric fitness of the assessment tool, recognize items that might be required to be revised or removed, and support a culture of evidence-based evaluation practices atthe primary level of education. The findings provide a helpful model for educators who want to evaluate their own assessment tools. They also offer useful guidance for test designers.

Objectives

1. To calculate and analyze the Difficulty Index (P) and Discriminating Power (D) for each item in a fifth-grade standardized multiple-choice examination.
2. To evaluate the overall psychometric quality of the examination based on item analysis results and provide specific recommendations for item revision or retention.

LITERATURE REVIEW

The Role and Evolution of Educational Assessment: Black and William (2018) assert that educational assessment has developed from being merely summative and selection-oriented to becoming a sophisticated instrument that is vital to educational development. According to DeMars (2018), formative assessment—often referred to as assessment for learning—uses evaluation to provide targeted feedback and regulate instructional strategies in real-time. Conversely, learning assessment, or summative assessment, assesses student achievement at the end of the process (Yu et al., 2024). When properly evaluated, standardized tests can offer formative insights even if they are commonly utilized for summative objectives (Crocker & Algina, 2008). The 5^{th} grade is the most important phase in education, where major reading, numerical, and scientific reasoning skills are consolidated. At this stage, valid and reliable evaluations are crucial for assessing a candidate's aptitude for advanced studies (Ghazali et al., 2023; COnsolata, 2024).

Classical Test Theory (CTT) and Its Foundational Assumptions: CTT serves as the primary milestone in traditional psychometric item analysis. The theory asserts that the true score (T) and the error score (E) give the observed test score (X). The true score (T) determines the examinee's real test score, and the error score (E) represents the random error in the measurement: X=T+E (Embretson & Reise, 2025). The main aim of the psychometric analysis is to refine the dependability, validity, consistency, and reliability of the conclusions drawn from observed test scores (T) while reducing the errors. Since CTT statistics are

sample-dependent, the ability distribution of the target group undergoing the assessment significantly influences both item difficulty and discrimination metrics (Pastore, 2023). Though there are many other complex models like IRT (Item Response Theory), CCT is still utilized worldwide. It is due to its spontaneous simplicity and suitability for school- and classroom-level evaluation (Bennett, 2015; Skaggs, 2021).

Item Difficulty: Conceptualization and Interpretation: The difficulty index (P) is a useful measure that describes the easiness or difficulty of an item. It ranges from 0.0 to 1.0, where higher values show more ease of an item (Flake & Fried, 2020). To interpret the maximum difficulty of an item, there are many methods. In criterion-referenced tests, which are meant to measure mastery, the *P* with a higher value is preferred. While norm-referenced tests are designed to predict variance for discrimination, the *P* with a moderate difficulty value is considered to be ideal. As per the normally acknowledged guideline, the maximum difficulty value for MCQs is normally higher than the chance value (1/k); the "k" denotes the number of options provided. So, for an MCQ with 4 options, a *P* with 0.50 to 0.70 is considered an excellent item, while an item having a *P* above 0.90 is taken too easily. Similarly, an item with P=0.20 is considered too difficult (Wormeli, 2017). The very easy items with $P > 0.90$ may indicate a ceiling effect and very poor curriculum alignment, whereas very difficult items with $P < 0.10$ may suggest that the content is irrelevant, overly specific, or improperly taught (Sanz et al., 2026).

Item Discrimination: The Core of Validity: The discriminating power (D) is widely considered the most crucial item statistic for norm-referenced evaluation because it is directly related to the validity of the item (Álvarez et al., 2026). The D index compares performance between a high-scoring group (often the top 27%) and a low-scoring group (bottom 27%) because this ratio best strikes a compromise between stability and differentiation (Khan et al., 2025). D ranges from -1.0 to +1.0. Positive values show that the top group performed better than the bottom group. The normal interpretation criteria are as follows: $D > 0.40$ = Very Good; $0.30 \leq D < 0.40$ = Good; $0.20 \leq D < 0.30$ = Marginal/Acceptable (may need improvement); and $D < 0.20$ = Poor (should be eliminated or amended) (Hattie & Zierer, 2024). Negative discrimination is a serious flaw that implies the question is either misleading, miss-keyed, or assesses something inversely related to the test's objective when more seasoned students select the wrong answer (Matazu & Juliu, 2021).

The Interplay between Difficulty and Discrimination: Discrimination and difficulty are interlinked. Due to their highest response difference, items with moderate difficulty ($P \sim 0.50$) suggest the maximum likelihood for greater discrimination (Fulcher & Harding, 2021). Logically, the very easy and the very difficult items have negligible discrimination, as nearly everyone responds to them correctly or wrongly, respectively. A moderately difficult item is, however, rationally connected to the knowledge or skill that the test is projected to evaluate in order to be a worthy discriminator (Cheung et al., 2024). A more responsive depiction can be gotten by examining both indices collectively: a difficult item having excellent discrimination may be measuring advanced, effective evidence, while an easier item with little discrimination may be a bad distractor (Luecht, 2025; Levy-Feldman, 2025).

Applications and Challenges in Primary Settings: Utilizing psychometric analysis in the elementary schooling has many complications, because at elementary stage, suitable comprehension levels, language, and cognitive necessities are very essential (Song et al., 2025). Finch et al. (2025) assert that beginners may be more vulnerable to test-wiseness artifacts and unclear item layouts. According to Gitomer (2026), psychometric exercises and time are mostly very rare for educators. The findings show that the psychometric analysis can

meaningfully advance the instructive usability, the consistency and the evidence-based teaching attitude while shifting evaluation from mere instinct to evidence-based judgment (Tushar & Sooraksa, 2023).

Beyond CTT: Item Response Theory and Computerized Adaptation: The Item Response Theory (IRT) has many benefits, especially for large-scale standardized tests and CAT (computerized adaptive testing), though CTT is more fit for many applications. IRT is more flexible while modeling item difficulty, item discrimination, and guessing because it is independent of the specific sample under evaluation (Pedaste et al., 2021). IRT needs more sophisticated software and a bigger sample size. IRT is not much applied for regular classroom application (Choi & McClenen, 2020). Now the gap is filling because of the introduction of the digitally designed evaluation tools that offer schools automatic item analysis and IRT standardization (Hayes et al., 2023).

Though these approaches are well documented in literature and are generally utilized in large-scale and high-stakes contexts, at the primary level there is still a perceptible gap while reporting and applying item analysis methods for both teacher-made and school-level standardized tests (Kumar et al., 2021). Most of the literature has focused on either university entrance examinations or national-level evaluation. Little literature is available on approachable and comprehensive findings regarding the item analysis of the subject of a particular class (Rock et al., 2026). It is also evident form literature that item difficulty and item discrimination have been treated separately, having less focus on the combined effect to provide useful and item-level findings for the test revision in everyday educational settings, with a minimum sample size and minimal resources (Hattie & Zierer, 2023).

While providing a detailed and open analysis of the 5^{th} class examination, the current study pursues filling the gap while describing the practical application of item analysis not only for assessment but also for the development of the testing tools, which have an effect on everyday teaching and instruction, especially in primary schools. The current study also focuses on the combined effect of the item difficulty and discrimination power, which was neglected in previous studies.

METHODOLOGY

The current study, while adopting a quantitative approach, has utilized the exam score data that was already available. The sample consists of fifth-class students who took the standardized test. The population consisted of 124 students. The sample population, which consisted of 62 students, was divided into two subgroups, high achievers and low achievers, as per the conventional CTT practices. Both groups usually compose 27% of high performers and 27% of lower performers, based on total test score (Rock et al., 2026). Both the Upper Group (H) and the Lower Group (L) separately totaled the number of correct answers for each of the fifteen MCQs.

Reliability of the Tool

Table No.1 Reliability Statistics

Cronbach's Alpha	No of Items
0.989	15

Table No. 1 presents a reliability value of 0.989 for the tool, indicating that the instrument utilized demonstrated excellent reliability (Ghazali et al., 2023).

Data Analysis

Two key indices (P and D) were calculated for every item:

Difficulty Index (P): P = (H + L) / N.

Here N shows the number of total students, both in the upper group and lower group. The implied N per item from the sum of correct responses column is 62 (e.g., for Item 1: 30+21=51, and 51/0.82 ≈ 62). This value was utilized for all succeeding calculations for reliability and uniformity.

Discriminating Power (D): D = (H - L) / (N/2).

The formula given standardizes the variance between the groups by the number of learners in one group, i.e., half of the whole evaluated sample.

Each item was classified while utilizing the established standards (Ghazali et al., 2023). For Difficulty: $P > 0.70$ = Easy; $0.30 \leq P \leq 0.70$ = Moderately Difficult; $P < 0.30$ = Difficult. For discrimination: $D \geq 0.40$ = Very Good; $0.30 \leq D < 0.40$ = Good; $0.20 \leq D < 0.30$ = Marginal/Acceptable; $D < 0.20$ = Poor.

RESULTS

Table No. 2 Showing Results Regarding Difficulty Index

Item No	Upper Group	Lower Group	Sum of Correct Reponses = (H+L)	Difficulty Index = (H+L)/N	Decision of Each Item
1	30	21	51	0.82	Easy Item
2	29	12	41	0.66	Moderately Difficult to Easy
3	14	6	20	0.32	Moderately Difficult to Easy
4	17	8	25	0.40	Moderately Difficult to Easy
5	30	15	45	0.73	Easy Item
6	28	7	35	0.56	Moderately Difficult to Easy
7	28	8	36	0.58	Moderately Difficult to Easy
8	20	3	23	0.37	Moderately Difficult to Easy
9	16	8	24	0.39	Moderately Difficult to Easy
10	22	6	28	0.45	Moderately Difficult to Easy
11	21	4	25	0.40	Moderately Difficult to Easy
12	22	2	24	0.39	Moderately Difficult to Easy
13	15	6	21	0.34	Moderately Difficult to Easy
14	24	5	29	0.47	Moderately Difficult to Easy
15	31	17	48	0.77	Easy Item

Table No. 2 describes the difficulty index (P) for every item. The analysis identified a variance in difficulty index (P).

It is evident from table no. 2 that three items: 1,5, and 15, having *P* values of 0.82, 0.73, and 0.77, respectively, have been answered correctly by the majority of students in both performing groups and thus declared as the easy ones. Similarly, it has also been observed that twelve items: 2, 3, 4,6, 7, 8, 9, 10, 11, 12, 13, and 14,with P values ($0.30 \leq P \leq 0.70$), are placed in the "moderately difficult" category, and declared as the ideal ones items. It is also obvious from table no. 2 that there was no item with $P < 0.30$.

General Test Difficulty: Throughan average *P* (averaging across questions) possibly in the 0.50-0.55 range, which is very good for making score variance, the occurrence of moderately difficult items designates the test was normally well-targeted at the cohort's skill level.

Table No. 3 Showing Results Regarding Discriminating Power

Item No	Upper Group	Lower Group	Difference of Correct Reponses = (H-L)	Discriminating Index = (H-L)/N/2	Decision of Each Item
1	30	21	9	0.29	Marginal/Acceptable Discriminator (Subject to Improve)
2	29	12	17	0.55	Very Good Discriminator
3	14	6	8	0.26	Marginal/Acceptable Discriminator (Subject to Improve)
4	17	8	9	0.29	Marginal/Acceptable Discriminator (Subject to Improve)
5	30	15	15	0.48	Very Good Discriminator
6	28	7	21	0.68	Very Good Discriminator
7	28	8	20	0.65	Very Good Discriminator
8	20	3	17	0.55	Very Good Discriminator
9	16	8	8	0.26	Marginal/Acceptable Discriminator (Subject to Improve)
10	22	6	16	0.52	Very Good Discriminator
11	21	4	17	0.55	Very Good Discriminator
12	22	2	20	0.65	Very Good Discriminator
13	15	6	9	0.29	Marginal/Acceptable Discriminator (Subject to Improve)
14	24	5	19	0.61	Very Good Discriminator
15	31	17	14	0.45	Very Good Discriminator

It is obvious from table no. 3 that ten out of fifteen items having $D \geq 0.40$, exhibited very good to exceptional discrimination power (2, 5, 6, 7, 8, 10, 11, 12, 14, and 15). Similarly, items no. 6, 7, and 12 having high D values of 0.68, 0.65, and 0.65, respectively, suggested a very good capacity of distinguishing between high-group performers and low-group performers.

It is also evident from table no. 3 that items no. 1, 3, 4, 9, and 13 were having D values from 0.20-0.30, categorized as acceptable/marginal discriminators. According to the research, these are "subject to improvement."

The test validity was found to be positive because no items showed poor ($D < 0.20$) or negative discrimination.

Integrated Analysis of Difficulty and Discrimination: Deeper understanding can be gained by comparing the two indices:

The analysis reveals that item no. 1, with *D*=0.29 and easy *P*=0.82, is declared as a marginal discriminator. The majority of the students, regardless of ability, answered it correctly. Its high difficulty index reduces the possibility of discrimination. It can be testing an idea that the entire cohort has mastered.

Likewise, four items, i.e., 3, 4, 9, and 13 showed minor discrimination, but highly challenging (*P* between 0.32 and 0.40). This inconsistency raises the probability of questions like unclear language, challenging but unrelated distractions, or assessing information unrelated to the overall concept of tests.

Similarly, item nos. 2, 5, 6, 7, 8, 10, 11, 12, 14, and 15 have been regarded as the strongest elements of the test. The majority of the discriminators are excellent and challenging. Item 5 (*P*=0.82), though easy, is still a very good discriminator (D=0.48), which is a required characteristic for an easy item. It still favors the high-performing group.

DISCUSSION

It is obvious from the results that the psychometric quality of the 5th class's standardized test is usually very good and positive, but still there are several items that require improvement. The majority of moderately difficult question items (12 out of 15) are consistent with norm-referenced tests, which are meant to yield a score distribution (Pastore, 2023). This suggests that the curriculum and proficiency level of the fifth-grade populace were magnificently targeted by the test designers.

The greatest promising outcome is the excellent discriminating capability of 10 items. The overall mastery of the content of the high-performing group constantly beats the lower-performing group on these items. The high discrimination indexes designate that the assessment is reliably assessing the intended academic concept (Flake & Fried, 2020). Items having a D value less than 0.60, like 6, 7, and 12, are particularly advantageous and must be retained as model items for forthcoming tests. However, a keen assessment of those five items, which fall into the marginal/acceptable discriminators categories, is necessary. Sanz et al. (2026) asserts that those items which have a D value in between 0.20 and 0.30 are usually regarded as doubtful and must be revised. Item No. 1 is an example of a psychometric phenomenon. Very easy or very difficult items cannot show variance in answers, which in turn limits the discrimination ability of the particular item(s) (Cheung et al., 2024). Such items might be included in a criterion-referenced test that verifies near-universal mastery, but they don't add much to a norm-referenced instrument's reliability and might be eliminated to make the test shorter without sacrificing measurement accuracy.

The moderately challenging items with low discrimination (3, 4, 9, and 13) are more problematic. Although their acceptable difficulty indicates that they are pitched at the proper level, their poor discrimination implies that there may be problems that are not obvious in the quantitative data alone. Qualitative diagnostic review is necessary for these items. Possible problems include:

1. Ambiguous Stem or Options: Even brilliant students may find the question or answer options unclear (Luecht, 2025).
2. Construct-Irrelevant Variance: Rather than testing the intended subject-matter knowledge, the item may inadvertently assess reading comprehension, cultural knowledge, or test-wiseness (Levy-Feldman, 2025).

The current study provides a basis for how classical item analysis can be used in real-world situations. Comparatively simple calculations produce a clear picture of the test's advantages and disadvantages. It transforms test evaluation into an actionable, item-by-item diagnostic, going beyond simple face validity or overall reliability coefficients. Adopting such a regular

practice will help fifth-grade teachers and assessment coordinators ensure that exams, used to make critical choices about students' development, act as learning tools themselves, continuously being assessed and improved based on data.

CONCLUSION

The current study performed a psychometric analysis of 15 items of the 5th grade standardized MCQs test to describe their difficulty indices (P) and discriminating power (D). The findings indicate that the testing tool is well-designed. Most of the items exhibited very good discrimination and a moderately difficult nature. This shows the basic validity of the test for distinguishing high achievers and low achievers in the assessed domain.

The current study also identified several items that needed to be addressed psychometrically. The four moderately difficult items (3, 4, 9, and 13) and one very easy item (item no. 1) exhibited very limited discrimination power (D). Such things increase the chances to improve the capability of the assessing tool. To ascertain and address any possible issues with paraphrasing, distracting quality, and content alignment, it is suggested that those five MCQ items may be subjected to inclusive qualitative evaluation by some content experts. The issues identified must be re-piloted once revised. The remaining ten items, which possess excellent discriminating power (D), provide a concrete base for future testing practices. Those can be utilized as models for generating new testing items.

Recommendations

The psychometric analysis of the items has determined a clear path to improve the inspected evaluation, as it is sound psychometrically. This designates a basic rule of educational measurement and evaluation. Effective assessments are not just carried out but rather designed, evaluated, and developed through repetitive and evidence-based procedures. To ensure the assessment process and procedures are impartial, precise, and correct and student-centered, it is important to take such analytical techniques into consideration in the evaluation process at the primary level of education.

The overall computed reliability coefficient and the distractor analysis of those five items, along with a comparison of the performance of the students on the revised items of the test, could be part of the forthcoming investigation.

Acknowledgment: The authors express sincere gratitude to the head teachers of primary schools (boys) for providing students' results data.

Conflict of Interests: The authors declare that no competing interests exist.

Authors' Contribution: The article is purely for research purposes. All three authors equally contributed to data collection, review of literature, data analysis, and manuscript preparation.

Funding Information: This research received no specific grant from any funding agency in the government, commercial, or not-for-profit sector.

REFERENCES

Álvarez, J. S., He, Q., Guenole, N., & D'Urso, D. (2026). Using Artificial Intelligence in Test Construction: A Practical Guide. Psicothema, 38(1), 1-12. https://www.doi.org/10.70478/psicothema.2026.38.01

Bandalos, D. L. (2018). Measurement Theory and Applications for the Social Sciences. Guilford Publications.

Bennett, R. E. (2015). The Changing Nature of Educational Assessment. Review of Research in Education, 39(1), 370-407. https://doi.org/10.3102/0091732X14554179

Black, P., & Wiliam, D. (2018). Classroom Assessment and Pedagogy. Assessment in Education, 25(3), 1-29. https://www.doi.org/10.1080/0969594X.2018.1441807

Brookhart, S., & Nitko, A. J. (2019). Educational Assessment of Students. Pearson.

Brown, G. T. (2019). Is Assessment for Learning Really Assessment?. In Frontiers in Education (Vol. 4, p. 64). Frontiers Media SA. https://doi.org/10.3389/feduc.2019.00064

Cheng, L., & Fox, J. (2017). Assessment in the Language Classroom: Teachers Supporting Student Learning. Bloomsbury Publishing.

Cheung, G. W., Cooper-Thomas, H. D., Lau, R. S., & Wang, L. C. (2024). Reporting Reliability, Convergent and Discriminant Validity with Structural Equation Modeling: A Review and Best-Practice Recommendations. Asia Pacific Journal of Management, 41(2), 745-783. https://doi.org/10.1007/s10490-023-09871-y

Choi, Y., & McClenen, C. (2020). Development of Adaptive Formative Assessment System Using Computerized Adaptive Testing and Dynamic Bayesian Networks. Applied Sciences, 10(22), 8196. https://doi.org/10.3390/app10228196

COnsolata, S. K. (2024). Exploring the Influence of Self-Efficacy on Competency-Based Teaching and Learning Outcomes in Grade Seven Junior Secondary School in Trans-Nzoia County, Kenya.1-98. https://ecommons.aku.edu/etd_tz_ied_m-ed/493

Crocker, L. M., & Algina, J. (2008). Introduction to Classical and Modern Test Theory. Cengage Learning.

DeMars, C. E. (2018). Classical Test Theory and Item Response Theory. The Wiley Handbook of Psychometric Testing: A Multidisciplinary Reference on Survey, Scale and Test Development, 49-73. https://doi.org/10.1002/9781118489772.ch2

Embretson, S. E., & Reise, S. P. (2025). Item Response Theory: Foundations for Psychologists and Social Scientists. Routledge. https://www.doi.org/10.4324/9781315726557

Finch, W. H., French, B. F., Immekus, J. C., & Dai, S. (2025). Educational and Psychological Measurement. Routledge. https://www.doi.org/10.4324/9781003439769

Fulcher, G., & Harding, L. (Eds.). (2021). The Routledge Handbook of Language Testing. Taylor & Francis Group. https://www.doi.org/10.4324/9781003220756

Furr, R. M. (2021). Psychometrics: An Introduction. SAGE Publications.

Ghazali, M. T. B., Ghani, M. A., Rahman, S. A. A., Afthanorhan, W. M. A. B. W., Fauzi, M. A., & Wider, W. (2023). Validation of a Scale for the Measurement of Employee Competency in Relation to Succession Planning amongst Administrators in Higher

Education Institutions. International Journal of Professional Business Review, 8(4), 5. https://doi.org/10.26668/businessreview/2023.v8i4.1233

Gitomer, D. H. (2026). Evidence Centered Design: The Contribution of Bob Mislevy in Advancing Validity Theory. Educational Measurement Issues and Practice, 45(1), 1-9. https://www.doi.org/10.1111/emip.70033

Hattie, J., & Zierer, K. (2024). 10 Mindframes for Visible Learning: Teaching for Success. Routledge. https://www.doi.org/10.4324/9781003430124

Hayes, N., O'Toole, L., & Halpenny, A. M. (2023). Introducing Bronfenbrenner: A Guide for Practitioners and Students in Early Years Education. Routledge. https://www.doi.org/10.4324/9781003247760

Heritage, M. (2021). Formative Assessment: Making it Happen in the Classroom. SAGE Publications. https://doi.org/10.4135/9781071813706

Khan, H. F., Qayyum, S., Beenish, H., Khan, R. A., Iltaf, S., & Faysal, L. R. (2025). Determining the Alignment of Assessment Items with Curriculum Goals Through Document Analysis by Addressing Identified Item Flaws. BMC Medical Education, 25(1), 200. https://doi.org/10.1186/s12909-025-06736-4

Kumar, D., Jaipurkar, R., Shekhar, A., Sikri, G., & Srinivas, V. (2021). Item Analysis of Multiple Choice Questions: A Quality Assurance Test for an Assessment Tool. Medical Journal Armed Forces India, 77, S85-S89. https://doi.org/10.1016/j.mjafi.2020.11.007

Levy-Feldman, I. (2025). The Role of Assessment in Improving Education and Promoting Educational Equity. Education Sciences, 15(2), 224. https://doi.org/10.3390/educsci15020224

Luecht, R. M. (2025). Assessment Engineering in Test Design: Methods and Applications. Routledge. https://www.doi.org/10.4324/9781003451051

Matazu, S. S. A., & Julius, E. (2021). Item Analysis: A Veritable Tool for Effective Assessment in Teaching and Learning. Journal of Education and Practice. https://www.doi.org/10.7176/JEP/12-21-04

Pastore, S. (2023, July). Teacher Assessment Literacy: A Systematic Review. In Frontiers in Education (Vol. 8, p. 1217167). Frontiers Media SA. https://www.doi.org/10.3389/feduc.2023.1217167

Pedaste, M., Baucal, A., & Reisenbuk, E. (2021). Towards a Science Inquiry Test in Primary Education: Development of Items and Scales. International Journal of STEM Education, 8(1), 19. https://doi.org/10.1186/s40594-021-00278-z

Rock, M. L., Billingsley, B., Dieker, L., & Leko, M. (Eds.). (2026). Transforming the Special Education Workforce: Research and Complex Systems Perspectives. American Educational Research Association. https://doi.org/10.2307/jj.42063210

Sanz, S., García, C., & Olmos, R. (2026). Use of Crossed Random-Effects Models to Assess Multiple-Choice Items: An Experimental Study. The Spanish Journal of Psychology, 29(8), 1-10. https://doi.org/10.1017/SJP.2026.10022

Skaggs, G. (2021). Test Development and Validation. SAGE Publications.

Song, Y., Du, J., & Zheng, Q. (2025). Automatic Item Generation for Educational Assessments: A Systematic Literature Review. Interactive Learning Environments, 33(9), 5386-5405. https://doi.org/10.1080/10494820.2025.2482588

Tushar, H., & Sooraksa, N. (2023). Global Employability Skills in the 21st Century Workplace: A Semi-Systematic Literature Review. Heliyon, 9(11). https://www.doi.org/10.1016/j.heliyon.2023.e21023

Wormeli, R. (2017). Fair isn't Always Equal: Assessment and Grading in the Differentiated Classroom. Stenhouse Publishers. https://www.doi.org/10.4324/9781032681122

Yu, H. C., Shih, Y. A., Law, K. M., Hsieh, K., Cheng, Y. C., Ho, H. C., ... & Fan, Y. C. (2024, August). Enhancing Distractor Generation for Multiple-Choice Questions with Retrieval Augmented Pretraining and Knowledge Graph Integration. In Findings of the Association for Computational Linguistics: ACL 2024 (pp. 11019-11029). https://www.doi.org/10.18653/v1/2024.findings-acl.655