The Eurasia Proceedings of Educational & Social Sciences (EPESS), 2024

Volume 36, Pages 70-77

ICRESS 2024: International Conference on Research in Education and Social Sciences

Evaluating the Level of Difficulty of University Statistical Course Assessments: A Modern Perspective through Item Analysis

Sunil Prakash Pillai
University of Technology and Applied Sciences

Manitha Rijo
University of Technology and Applied Sciences

Abstract: University statistics courses serve as critical platforms for developing learners' mathematical thinking, problem-solving skills, and analytical abilities. To gauge the effectiveness of these courses and to provide learners with a fair and accurate assessment of their understanding, it is imperative to design assessments that appropriately balance the level of difficulty. Striking the right balance between challenging questions that stimulate critical thinking and accessible questions that gauge fundamental knowledge is a complex endeavor. In this context, item analysis, a data-driven technique for assessing the performance of individual test items, emerges as a valuable tool to ensure the quality and fairness of statistics assessments in higher education. In this study, we seek to explore the contemporary applications of item analysis in the evaluation of the level of difficulty in university statistics course assessments. With the rapidly evolving landscape of educational research and the advent of advanced statistical methodologies, this study aims to provide a modern perspective on item analysis techniques that can empower educators to create more effective statistics assessments. In this paper, we use the facility index and discrimination index to compare the assumed difficulty level and expected difficulty level of questions in the final exam of an advanced diploma course at the University of Technology and Applied Sciences (UTAS), Nizwa, Sultanate of Oman.

Keywords: Assessment validity, Item analysis, Tertiary education, Course assessments,

Introduction

Course assessments play a central role in evaluating learners' understanding of concepts, and ability to apply statistical techniques and skills in various contexts. This survey explores the academic processes involved in designing assessments for statistical courses in higher education along with the challenges faced by educators in this domain. An assessment encompasses a comprehensive process aimed at gathering data to facilitate decision-making regarding learners, curricula, programs, schools, and educational policies (Popham, 2013). When referring to "assessing a student's competence", it denotes the acquisition of information to determine the extent to which the learner has accomplished the intended learning outcomes (Shepard, 2016). Many assessment techniques may be deployed which are available for gathering such information. These include both formal and informal methods such as observations, paper-and-pencil tests, performance assessment tasks, and their corresponding marking schemes (Brookhart & Nitko,2019).

Assessment techniques include a wide variety of evaluative measures, including but not limited to a learner's performance on homework assignments, practical work, projects, and viva voce sessions (Wiggins & McTighe, 2013). Further to this technique, an analysis of the learners' academic records provides another valuable source of assessment data (Stiggins, 2014). These diverse methods collectively contribute to a holistic understanding of learners' learning and progress thereby aiding educators in making informed instructional decisions (Black & William, 2018).

© 2024 Published by ISRES Publishing: www.isres.org

Assessment Development

In an educational environment, the assessment development process is completed with the following levels,

a) *Designing Assessment Templates* – In this step, the assessment design process is initiated by outlining the content, format, and structure of the assessment. (American Educational Research Association, Americal Psychological Association & National Council on Measurement in Education, 2014).
b) *Developing Assessment Items or Tasks* – The second step involves creating specific assessment items or tasks aligned with the learning objectives and content standards. These include Multiple-Choice Questions (MCQs), Short Answer type Questions (SAQs), essays, performance tasks, or other types of assessments (Brookhart, 2013)
c) *Piloting and Refining* – An essential step in the process is to administer the assessment of the target population, before administering the test to the entire population. Assessment items may be revised or deleted to improve the quality of the assessment (Brookhart & Nitko, 2019).
d) *Setting Standards and Criteria* – Establishing performance standards or criteria defines what constitutes acceptable performance on the assessment. This involves determining proficiency levels that delineate different levels of performance (Baker et al., 2010).
e) *Scoring and Analysis* – Scoring involves applying predetermined scoring rubrics or criteria to evaluate learner responses. After scoring, data analysis techniques are employed to interpret assessment results, including examining item difficulty, discrimination, and other psychometric properties (Brennan, 2006).
f) *Interpreting and Using results* – Finally, assessment results are interpreted to make educational decisions at various levels, such as informing instructional planning, evaluating program effectiveness, or making high-stakes decisions about learner achievement (Stiggins, 2012).

Significance of Assessment in Education

Assessment in tertiary education serves as a fundamental component of the learning process playing a crucial role in evaluating learner achievement, promoting deep learning, and driving educational improvement (Boud & Falchikov, 2013). Through assessment, educators can gauge learners' understanding of course content, critical thinking abilities, and mastery of key concepts (Freeman et al., 2014). Moreover, assessment provides valuable feedback to both learners and instructors, informing instructional decisions and guiding future learning activities (Black & William, 2018). By providing timely and constructive feedback, formative assessment empowers learners to identify areas for improvement, set learning goals, and stake ownership of their learning process (Hattie & Timperley, 2017). The feedback received post-assessment enables learners to monitor their progress and adapt their study strategies accordingly (Gibbs & Simpson, 2013).

Higher education institutions rely on assessment to ensure accountability and quality assurance to their stakeholders (Alreck & Settle, 2014). Institutions must provide assessment data to accrediting bodies and educational authorities to evaluate program effectiveness, monitor learner achievement, and uphold standards of academic excellence (Shepard, 2016). Assessments play a vital role in evaluating learning progress, identifying areas of strengths and weakness, and informing instructional interventions to support learner growth and achievement (Brookhart, 2013).

Challenges in Assessment Preparation

a) *Validity and Reliability* - Educators face challenges while ensuring the validity and reliability of statistical assessments. Assessments that pass the validity test accurately measure the intended learning outcomes, while reliable assessments produce consistent results each time they are administered. Designing assessments that strike a balance between validity and reliability requires careful attention to assessment item quality, test construction, and psychometric properties (Scheaffer et al, 2016).

b) *Addressing different backgrounds* – Learners enrolled in courses will invariably belong to different academic backgrounds especially their level of mathematical preparation, statistical knowledge, and quantitative skills. Educators need to accommodate this diversity while preparing assessment materials. This may involve providing additional support resources for learning and offering opportunities for remedial assessment and advanced learning (Pfannkuch et al, 2017).

Item Analysis - Role in Assessment Design

Item analysis is an audit process used to evaluate the quality and effectiveness of individual test items or questions within an assessment. It involves analyzing various item statistics such as difficulty index, discrimination index, and effectiveness to identify items that need revision or removal to improve the overall reliability and validity of the assessment.

The validity of an assessment tool is the extent to which it measures what it was designed to measure, without contamination from other characteristics. For example, a test of reading comprehension should not require mathematical ability. The research on ensuring the validity of assessment items focuses on content validity, construct validity, and criterion-related validity.

a. *Content validity* refers to the extent to which an assessment item adequately represents the domain of interest and aligns with the intended learning objectives. Researchers (Kane, 2013; Downing,2019) have explored methods for assessing and enhancing content validity through expert judgment, alignment analyses, and curriculum mapping. Content validity is typically established through a systematic process. Item Analysis is used here to identify whether test items cover the full range of topics or skills specified in the learning objective. For instance, if a mathematics test is designed to measure the learner’s knowledge of mathematical operations such as addition and subtraction but only includes multiplication and division, the validity of the assessment would be questionable.

The content validity process includes the following steps

i. *Defining Learning Objectives* – This is the first step in the process. Here the learning objectives, that the assessment item must measure, are clearly defined. These must be aligned with the course curriculum and educational standards.
ii. *Item Development* – The second step involves aligning assessment items with the defined learning objectives. This may involve drafting assessment items that may include multiple choice questions, short answer questions, or tasks that cover the scope of the course in-depth and breadth.
iii. *Domain Experts’ Reviews* – The third step requires that the Assessment items are reviewed by Subject Matter Experts, to ensure that they are relevant, representative, and sufficiently comprehensive in covering the content domain. This expert judgment is important for validating the content validity of the assessment.
iv. *Pilot Testing* – The fourth step requires that all the assessment items are pilot-tested with a sample of learners, to gather feedback on item clarity, appropriateness, and relevance. Any necessary revisions are made based on these pilot test results.
v. *Alignment Analysis* – The final step involves assessing the degree to which the assessment items align with the defined learning objectives and content standards. This analysis provides empirical evidence of content validity.

For instance, if a learning objective of a statistical course is “Perform a Chi-square test”, then the corresponding assessment item must be able to evaluate the learners' ability to comprehend the question item and provide the correct solution.

b. *Construct Validity* refers to the assessment items accurately measuring the underlying construct or theoretical concept. For instance, if a test is designed to measure intelligence, but only measures memorization skills then it lacks construct validity (Messick, 1989). In a statistical course, refers to the degree to which the assessment measures statistical knowledge and skills.

For example, when designing an assessment for the “Hypotheses Testing” course, the instructor will include questions, that require learners to correctly identify null and alternate hypotheses, choose the appropriate test statistic, and interpret the results. Similarly, for the “Regression Analysis” course, the instructor will include problems that require learners to perform simple and multiple regression anlaysis, interpret regression coefficients and asseess the goodness-of-fit of the model.

Statistical Methods for Establishing the Difficulty Level of University Educational Assessments

This study delves into the nuanced evaluation of difficulty levels within university statistics course assessments, employing a modern perspective through rigorous item analysis. The ever-evolving landscape of education necessitates a re-evaluation of traditional assessment methods to ensure they align with contemporary pedagogical approaches. Through a systematic examination of individual test items, this research aims to provide valuable insights into the intricacies of difficulty levels, shedding light on the effectiveness and relevance of current assessment practices in the field of university-level statistics education.

The methodology integrates statistical techniques to analyze the performance of learners across various question types, exploring patterns of difficulty and discrimination. By employing this comprehensive approach, the study seeks to identify specific content areas that may pose challenges to learners, thus enabling educators to make informed adjustments to teaching methodologies and assessment designs.

The research also considers the impact of technological advancements on assessment strategies, recognizing the potential of innovative tools and methodologies in enhancing the accuracy of difficulty assessments. The findings contribute to ongoing discussions on the continuous improvement of statistics education at the university level, addressing the evolving needs of both learners and educators.

Ultimately, this research endeavors to bridge the gap between traditional assessment practices and modern pedagogical demands, offering a fresh perspective on evaluating difficulty levels in university statistics course assessments. The insights garnered from this study are anticipated to inform future educational practices, promoting an adaptive and learner-centered approach to teaching and assessment in the realm of higher education statistics. Moodle offers a Statistics report for Multiple Choice questions (MCQs), (Moodle, 2023). This report provides a psychometric analysis of both the quiz as a whole and its individual questions. Users may choose to view the report online or download it in spreadsheet format for further examination.

The study aims to present a comprehensive method for evaluating difficulty levels in university short answer Questions (SAQs). The research focuses on relating Bloom’s taxonomy to the Item analysis to align the cognitive levels of Bloom's taxonomy with the difficulty and discrimination levels of the assessment items. The statistical analysis was carried out using MS Excel 2010 and SPSS ver 20.0, to analyze the data obtained from the assessment tool.

The examination of an assessment using item analysis offers valuable insights into its level of difficulty. In a research conducted by Kumar et al. (2021), item analysis was conducted on a total of 90 multiple-choice questions (MCQs) across three tests administered to 150 first-year Bachelor of Medicine and Bachelor of Surgery (MBBS) physiology learners. The research helped the instructors to identify good or ideal assessment items. These were to be included for future assessments or revision purposes. The study also identified those assessment items that needed to be revisited or improvised. The item Analysis process is aimed at examining the difficulty index (DIF I) and discrimination index (DI), along with assessing distractor effectiveness (DE).

Facility index (F) / Difficulty Index

The facility index measures the proportion of learners, who answered a particular question or assessment item, correctly (Brennan, 2006). The higher difficulty indices indicate easier items, while lower indices suggest more challenging items. In our study, we have implemented the Difficulty Index (P), calculated as per the statistical formulas in (Moodle, 2023), as shown below.

$$\text{Facility Index} = (X_{\text{Average}}) / X_{\text{Max}}$$

X_{Average} is the mean score obtained by all users attempting the test item and X_{Max} is the maximum score achievable for that test item.

Discrimination Index (D)

The discrimination index assesses how well an item differentiates between high and low performers in the overall test (Brown et al, 2013; Tavalol & Dennick, 2011). Positive values indicate good discrimination, negative values suggest poor discrimination, and values around zero imply the item is not effective in distinguishing between high performers and low performers. The discrimination index is calculated by correlating the score of each individual item with the total test score, excluding the item in question.

Evaluating Assessment Items based on a post-assessment approach

Experiment Design

In the study, 62 learners from two sections of a Statistics Course in the Advanced Diploma program, Fall Semester were administered, an assessment of 15 questions, 10 MCQ, and 5 SAQ. For the assessment preparation process, the instructor sets the assessment by mapping each assessment item to the specific Learning Outcomes and Design Rubrics. During the assessment period, learners attempt the assessment and submit it. The instructor evaluates the submitted assessments. The individual learners' scores are collated and published. The post-assessment step is crucial to this study. In this step, the assessments are analyzed using the Item Analysis process. For our research, Microsoft Excel software was used to analyze the collated scores. Bloom’s taxonomy was applied to categorize the questions according to the required cognitive levels. The University’s exam setting guidelines require that the difficulty levels of assessment items for the Advanced Diploma course should be distributed as follows: 30% easy, 40% moderately difficult, and 30% difficult. These guidelines have been revised for the new academic year, as 20% easy, 50% moderately difficult, and 30% difficult.

Post-assessment results showed that the failure rate for the course was greater than 30%. These results were further examined as a part of the routine University Quality Audit process. This process stipulates that if the number of failures or number of A grades exceeds 30% then the results will be considered incongruous. These results are then analyzed using the Item Analysis method, to compare the expected and actual difficulty levels. A report was submitted by the course coordinator, to the department for Quality Audit purposes. The methodology adopted for this research has been illustrated, is shown in Figure 1.

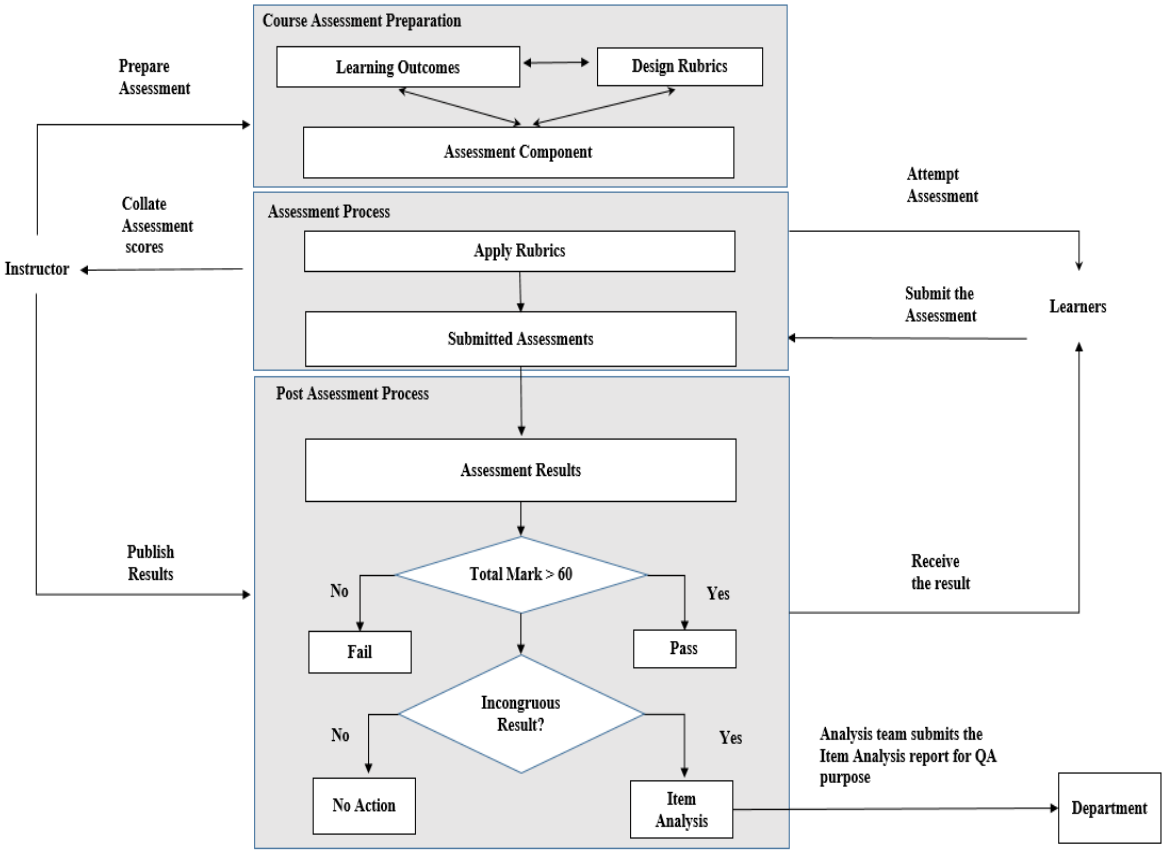

Figure 1. Item analysis process

Facility Index (F) - In Table 1, we can see that the facility index has been calculated. The expected difficulty level and actual difficulty level of these questions have also been identified. These questions have been labeled E, M or D, or to indicate Easy Questions, Moderate Difficulty Questions, or Difficult Questions.

Discrimination Index(D) - The Discrimination Index (D), refers to the correlation value between weighted scores on one of the questions and those on the rest of the test questions, except for that question. This classification scheme was adopted for interpreting the results In Table 3, scores 50 and above, indicate *very good* discrimination, scores between 30 and 50 indicate *adequate* or satisfactory discrimination, scores between

20 and 29, indicate *weak* discrimination, scores between 0 and 19, *very weak* discrimination and negative scores indicate these questions might be probably *invalid* and need to be verified.

Table 1. Questions and the corresponding facility index (F), expected difficulty & actual difficulty

Qn. No	Facility Index (F)	Expected Difficulty	Actual Difficulty
A1	0.75	E	E
A2	0.76	D	E
A3	0.51	M	M
A4	0.27	M	D
A5	0.59	E	M
A6	0.59	M	M
A7	0.40	D	M
A8	0.16	D	D
A9	0.19	D	D
A10	0.16	M	D
B1	0.58	E	M
B2	0.34	M	D
B3	0.49	M	M
B4	0.69	E	E
B5	0.52	D	M

Table 2. Questions and the corresponding Discrimination Index

Question No.	Discrimination Index
A1	0.23
A2	0.21
A3	0.32
A4	0.37
A5	0.47
A6	0.38
A7	0.27
A8	0.07
A9	0.01
A10	0.19
B1	0.52
B2	0.62
B3	0.56
B4	0.35
B5	0.43

Table 3. Classification of discrimination index

Index	Interpretation
0.5. and above	Very good discrimination
0.3- 0 .5	Adequate discrimination
0.20-0.29	Weak discrimination
0.0 -0.19	Very weak discrimination
-ve	Questions probably invalid

Result

a. Initially, it was assumed that the expected difficulty level would be 30% for easy questions, 40% for moderate questions, and 30% for difficult questions. However, the *actual difficulty level* was 20% for easy questions, 53% for moderate questions, and 27 % for difficult questions.
b. In the case of the SAQs, the discrimination index value is very good or adequate. However, in the case of multiple-choice questions, weak discrimination values were observed.

Very weak discrimination values were observed for questions A8 and A9, which were categorized as difficult questions. It was inferred that it was because the topic was taught in the last week of the Course Delivery Plan, and attendance was very poor. This led to learners guessing the answers to these questions.

c. The coefficient of skewness of the overall course score was 0.2, indicating a distribution close to normal.
d. The facility index table shows that questions A5 and B1 were expected to be easy but they turned out to be moderate for learners. These questions are categorized as remembering and understanding types. Furthermore, the discrimination analysis shows that questions B1, B2, and B3 have a discrimination index of 50 and above, indicating very good discrimination. Similarly, the short answer questions B4 and B5 with a discrimination index above 30, indicate adequate discrimination.

Scientific Ethics Declaration

The authors declare that the scientific ethical and legal responsibility of this article published in EPESS journal belongs to the authors.

Acknowledgments or Notes

* This article was presented as an oral presentation at the International Conference on Research in Education and Social Sciences (www.icress.net) held in Tashkent/Uzbekistan on August 22-25, 2024

* The authors would like to express their gratitude to the former Head of the Information Technology Department Dr. Duhai Al Shukaili and the present Head of the Information Technology Department Ms. Suad Al Riyami, UTAS Nizwa for their valuable support in supporting this research.

References

Alreck, P. L., & Settle, R. B. (2014). *The survey research handbook* (4th ed.). McGraw-Hill.

Black, P., & Wiliam, D. (2018). Classroom assessment and pedagogy. *Assessment in Education: Principles, Policy & Practice*, *25*(6), 551-575.

Jan -Elen, M.J., David –Merill, M., & Micheal- Spector, J. (2013). *Handbook of research on educational communications and technology* (4th ed., pp. 785-794). Springer.

Boud, D., & Falchikov, N. (2013). *Rethinking assessment in higher education: Learning for the longer term* (2nd ed.). Routledge.

Brookhart, S. M. (2013). *How to assess higher-order thinking skills in your classroom*. ASCD.

Brookhart, S. M., & Nitko, A. J. (2019). *Educational assessment of learners.* Upper Saddle River, NJ: Pearson.

Brown, G. A., Bull, J., & Pendlebury, M. (2013). *Assessing learner learning in higher education*. Routledge.

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases learner performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences, 111*(23), 8410-8415.

Gibbs, G., & Simpson, C. (2013). Conditions under which assessment supports learners' learning. *Learning and Teaching in Higher Education, 1*(1), 3-31.

Hattie, J., & Timperley, H. (2017). The power of feedback. *Review of Educational Research, 77*(1), 81-112.

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.

Moodle.(2023,February2). *Quizstatisticsreport*. Retrieved from docs.moodle.org/404/en/Quiz_statistics_report

Pfannkuch, M., Budgett, S., & Reading, C. (2017). Enabling all learners to succeed in learning statistics: A framework for quality learning and teaching. *International Association for Statistical Education Roundtable.*

Plake, B. S., & Wise, L. L. (2014). What is the role and importance of the revised AERA, APA, NCME standards for educational and psychological testing?. *Educational Measurement: Issues and Practice, 33*(4), 4-12.

Shepard, L. A. (2016). The role of assessment in a learning culture. *Educational Researcher, 29*(7), 4-14.

Stiggins, R. (2012). *An introduction to learner-involved assessment for learning* (6th ed.).Pearson.

Strauss, M. E., & Smith, G. T. (2009). Construct validity: Advances in theory and methodology. *Annual Review of Clinical Psychology*, *5*, 1-25.

Tavakol, M., & Dennick, R. (2011). Post-examination analysis of objective tests. *Medical Teacher*, *33*(6), 447-458.

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review, 67*(3), 223-248.

Author Information

Sunil Prakash Pillai
University of Technology and Applied Sciences
Nizwa, Oman
Contact e-mail: *sunil.prakash@utas.edu.om*

Manitha Rijo
University of Technology and Applied Sciences
Nizwa, Oman

To cite this article:

Pillai, S.P., & Rijo, M. (2024). Evaluating the level of difficulty of university statistical course assessments: A modern perspective through item analysis. *The Eurasia Proceedings of Educational and Social Sciences (EPESS), 36,* 70-77.