FULL-LENGTH ARTICLE

Quality Considerations in the Design of Teacher Licensure Tests in Higher Education: A Case from Novice Biology Teachers in Ethiopia

Adane Hailu Herut[1]* & Brinda Oogarah-Pratap[2]

[1]Center for Education Studies & Research, Dilla University, Dilla, Ethiopia
[2]Mauritius Institute of Education, Mauritius. E-mail: b.oogarah@mie.ac.mu
*Corresponding author: adaneh@du.edu.et

ABSTRACT

Evaluating the quality of teacher licensing tests is crucial for enhancing the licensing system as well as determining the overall teacher competence, especially when a written test serves as the sole criterion for granting or denying a license. This study, rooted in classical test theory principles, sought to offer evidence-based insights into the quality of test items designed for licensing biology teachers in the Ethiopian context. The analysis centred on crucial classical test theory metrics, including the difficulty index, discrimination index, validity, and distractor efficiency. To accomplish this, response sheets from the 100-item teacher licensing test were gathered from 311 candidates spanning eight public Ethiopian universities. The analysis honed in on factors such as the difficulty index, discrimination index, and distractor efficiency. Additionally, the internal consistency reliability was assessed using the Kuder-Richardson Formula 20 (KR20). The results of the item analysis revealed that the test items demonstrated a satisfactory level of difficulty and distractor efficiency, with a reasonable discrimination index. However, the internal consistency reliability for the test did not meet the desired standard. To improve the quality of future tests within the licensing system, the study suggests incorporating various quality assurance measures including adherence to standardized guidelines of indices. Additionally, further recommendations and implications were forwarded based on the key findings.

Keywords: Classical test theory; discrimination index; item analysis; item quality; teacher licensing test

INTRODUCTION

Many developing nations have undertaken major educational reforms to improve their education systems in recognition of the important contribution of quality education to national development (Mensah et al., 2020). One of the key focuses of the reforms has been on improving the quality of teachers through the introduction of teacher licensure examinations. There is ample evidence that the quality of any education system largely depends on the quality of its teachers, and passing the licensure examination is an effective indicator of teacher quality (Rockoff, 2004; Aaronson et al., 2007; Harris and Sass, 2007). Initial teacher licensure examinations are commonly taken by qualified novice teachers in the form of written tests. The tests are intended to assess subject knowledge and skills identified by panels of educators as essential for entry into the teaching profession (Mitchell & Barth, 2001). Standardized processes are used to score the tests to help make the distinction between those who have the desired level of competence to start teaching and those who do not (National Research Council, 2001; Mensah et al., 2020).

The government of Ethiopia recognized the importance of improving its teaching force and thus introduced an initial teacher licensure examination in line with the fourth and fifth issues of the Education Sector Development Program (Ministry of Education, 2010; 2015). As one of the structures of the education sector, the Directorate for Teacher and School Leader Licensing in the Ministry of Education is the only authorized body to regulate, design, and administer the examination in the form of a written test for novice and experienced teachers in the country. Individuals who have completed a first degree and a postgraduate diploma in teaching and aspire to embrace a teaching career are required to take the only written test. The graduates are issued a teaching license if they have scored a minimum of 70% on the written test.

Studying teacher licensing tests and evaluating their quality are vital to strengthening the licensing system because assessment is an integral aspect of teacher licensure (Braun, 2005; Rogers, et al., 2020). In Ethiopia, there is a considerable policy focus on the development of the teaching force (Ministry of Education, 1994). However, to date, there has been a lack of research conducted on the quality of the written teacher licensing tests. There is a need for studies that gather validity evidence to gauge the quality of the test items (National Research Council, 2001), more so when performance on a written test provides the sole basis for determining whether a teacher can be issued or denied a license (Talal et al., 2015).

Thus, a study was conducted to assess the quality of the test items for the initial licensing of biology teachers in Ethiopia. The biology licensing test was selected since there have been concerns about the quality of biology teachers contributing to the decline in students' performance in biology over recent years (Egun, 2016). The study aims at generating evidence-based information on whether the current test is appropriately designed for the identification of novice teachers with the desired level of competence. Such information would support the design of licensing examinations that are effective in making the distinction between those who have the desired level of competence to start teaching and those who do not, thereby contributing to the quality of biology teaching at a national level. The recommendations would be relevant to education authorities involved in the design of teacher licensure examinations for biology as well as other subject areas. Moreover, the study is intended to serve as a theoretical basis for guiding documents for research in the area of assessment of the quality of tests for teacher licensure.

METHODS AND MATERIALS

Theoretical Assumptions

There are a number of theories that underlie the concept of test item analysis. This study, however, used a classical test theory to ground its unit of analysis. Classical Test Theory (CTT) stands as a widely embraced psychometric framework utilized for assessing test quality and gauging individuals' attributes (Brown, 2010). Proponents of CTT underscore its simplicity and practicality, asserting its efficacy in furnishing valuable insights into item difficulty and discrimination (Kirwan, 2017). The notable contributions of Frederic M. Lord and Lee J. Cronbach have played a pivotal role in advancing CTT, emphasizing its credibility and consistency (Lord & Novick, 1968; Moss, 1992). CTT's significance lies in its capacity to facilitate meaningful score interpretation, thereby empowering informed decision-making in various domains such as education, employment, and clinical assessment (Raîche, 2005). Despite the existence of alternative theories, CTT maintains its influence owing to its well-established assumptions, practical utility, and unwavering support from advocates like Hambleton and Jones (1993).

Classical Test Theory (CTT) serves as a foundational framework, guided by premises that address key aspects of evidence-based test quality, such as item difficulty, discrimination, reliability, and distractor efficiency (Kirwan, 2017). According to Suciati, et al. (2020), CTT posits that tests should encompass items with diverse difficulty levels, ensuring a thorough evaluation of candidates across a wide spectrum of skills and knowledge. Item difficulty is quantified by calculating the ratio of test takers who answered an item correctly to the total number of test takers (Moore et al., 2023). In addition, CTT further assumes that test items should effectively discriminate between individuals with varying levels of knowledge or ability (Kirwan, 2017). Moreover, Reliability is a critical assumption in CTT, emphasizing the need for licensure tests to produce consistent and dependable results (Haladyna, 2004). Cronbach's Alpha (α), a widely accepted measure of reliability in CTT, is employed to assess internal consistency, taking into account the number of items and the variance of both item scores and total scores (Tavakol & Dennick, 2011). Furthermore, CTT posits that test distractors should effectively challenge test takers (Kirwan, 2017). Distractors also need to be plausible and likely to be chosen by individuals lacking the measured knowledge or skill, while avoiding incorrect and less plausible options (Moore, et al., 2023). These assumptions play a pivotal role in ensuring the validity and reliability of assessments for candidates seeking licensure. The theory underscores that, by adhering to these principles, tests can offer a comprehensive evaluation of individuals' abilities, thereby facilitating well-informed decisions regarding their suitability for teaching positions.

The Study Design

This study analysed test item responses for new biology teachers using quantitative data from the Ethiopian Ministry of Education's database. The test item analysis, which was informed by classical test theory and focused on key item properties mentioned in the previous sections. The teacher licensing test, comprising 100 multiple-choice items, was administered to 311 registered teacher candidates from eight public universities in Ethiopia. The answer sheets of the 311 teacher candidates were collected for test item analysis. Each item comprised 4 options (1 correct answer and 3 distractors), and each correct answer was valued at 1 mark. The scores of 311 examinees were encoded into an in-house score automation template produced using SPSS version 27. Then, descriptive statistics were computed to examine the scores of the examinees in terms of the mean, standard deviation, and associated counts and percentages.

Method of Data Analysis

Item analysis involved grouping and categorization of the difficulty index, discrimination index, and distractor efficiency. The internal consistency reliability was computed using the Kuder-Richardson Formula 20 (KR20) (Anselmi, et al., 2019). The items on the teacher licensing test were analysed in terms of the number of indices as described next. For the item analysis, teacher candidates’ scores were sorted in descending order. Eighty-four (27%) teacher candidates were defined as "high" scorers with a mean score of 75.07, and a further 84 (27%) were defined as "low" scorers with a mean score of 33. The scores of the remaining 143 (46%) teacher candidates were not used in the analysis. The high scorer group was labelled as "H", and the low scorer group was tabbed as "L".

a) ***Difficulty Index (p-value):*** the range is from 0% to 100%, or more typically, as a proportion, from 0.0 to 1.00. Item difficulty was calculated for each item using the following equation:

$$p = \frac{H_c+L_c}{H_n+L_n}$$

where H_r represents the number of correct answers in the high group, L_r represents the number of correct answers in the low group, H_n represents the number of answers in the high group and L_n the number of answers in the low group.

b) ***Discrimination Index (DI):*** the point-biserial relationship describes how well examinees performed on the item and their total test score. It is also referred to as the Point-Biserial correlation (PBS), the range of DI is from 0.0 to 1.00. The DI is calculated using the following formula:

$$DI = \frac{H_r - L_r}{H_n \ or \ L_n}$$

c) ***Distractor Efficiency (DE):*** Distractor efficiency is determined on the basis of the number of non-functional distractors (NFDs) in an item. The percent of each distractor was calculated by using the following formula:

$$\text{Percent of a distractor} = \frac{number\ of\ examinees\ of\ the\ distractor}{total\ number\ of\ examinees}\ x\ 100$$

d) ***Test Reliability (KR20):*** is the amount of measurement error associated with a test score, where a higher value represents a more reliable overall test score. Typically, the internal consistency reliability is calculated to indicate how highly the items correlate with one another. High reliability indicates that the items all measure the same general construct. The formula for KR20 for a test with *K* test items numbered *i*=1 to *K*, where p_i is the proportion of correct responses to test item *i*, and q_i is the proportion of incorrect responses to test item *i* (so that *pi* + *qi* = 1) is:

$$r = \frac{K}{K-1}\left[1 - \frac{\sum_{i=1}^{K} p_1 q_1}{\sigma_x^2}\right]$$

And the variance for the denominator is:

$$\sigma_x^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n}$$

where n is the total sample size, X_i is the score of individual examinees and X is the mean total score. The value of KR20 can range from 0 to 1, with numbers closer to 1 reflecting greater internal consistency, indicating that the items are all measuring the same general construct. The widely accepted cut-off value of KR20 is greater than or equal to 0.7 (Feldt, 2006).

RESULTS

Test Parameters

Table 1 provides a comprehensive overview of the key test parameters examined in the multiple-choice exam. These parameters encompass item difficulty, discrimination, and distractor efficiency. The average p-value of 0.53 suggests a moderate level of difficulty for the test items, with 53% of examinees correctly answering the questions, indicating a balanced level of challenge.

In terms of discrimination, the mean index of 0.29 indicates that questions moderately differentiated between higher and lower scoring groups, striking a fair balance in assessing the examinees' varying levels of proficiency. Both the difficulty and discrimination indices fall within acceptable ranges, reflecting a well-constructed and equitable exam.

The mean distractor efficiency of 76.98% is noteworthy, indicating that, on average, examinees selected distractors 26.02% of the time. This suggests that the incorrect options were plausible and appealing to those lacking knowledge, thereby serving their purpose effectively. This finding underscores the careful design of response options to challenge and assess the depth of understanding among examinees. Additionally, the relatively low standard deviations for difficulty and discrimination indices imply a commendable consistency in these aspects across the test items. This consistency contributes to the reliability of the exam results, ensuring that the assessment remains fair and unbiased.

Table 1: Summary of test parameters

Parameter	Mean	SD
Difficulty Index (p-value)	0.53	0.21
Discrimination Index (DI)	0.29	0.19
Distractor Efficiency (DE)	76.98	29.1

Item Difficulty Index (p-value)

Table 2 provides a comparative analysis between the anticipated and actual distribution of item difficulty indices on the licensure test. Adhering to the established guidelines outlined by Hotiu (2006), the recommended composition for the test comprises approximately 5% easy items, 20% with medium-low difficulty, 50% of medium difficulty, 20% of medium-hard difficulty, and 5% hard difficulty items. Nevertheless, the findings reveal notable deviations from these prescribed standards. Specifically, there is a significant overrepresentation of items categorized as either very easy (13% as opposed to the expected 5%) or very hard (19% as opposed to 5%). Conversely, the test exhibits a scarcity of items falling into the medium-hard difficulty category, accounting for only 7% instead of the expected 20%. This disparity indicates an imbalance in the distribution of difficulty levels across the items, with a higher-than-anticipated number of questions falling at the extremes of being either very easy or very difficult. Moreover, a total of 32 items deviate from the acceptable ranges, highlighting the necessity for revisions to align with best practice guidelines. Ensuring a more balanced representation of difficulty levels among the items is imperative to enhance the overall quality and reliability of the licensure test.

Table 2: Expected and actual values in terms of item difficulty

p-value	Meaning	Expected Nº of Items	Actual Status	Difference
0.8 – 1.00	Easy items	5%	13%	-8%
0.6 – 0.8	Medium-low difficulty	20%	17%	3%
0.4 – 0.6	Medium difficulty	50%	44%	6%
0.2 – 0.4	Medium-high items	20%	7%	13%
0.0 – 0.2	Hard items	5%	19%	-14%

Discrimination Index

Table 3 showcases the discriminative power of test items, gauged by their discrimination index values. It categorizes items into ranges indicating varying degrees of discrimination effectiveness: poor, marginal, good, or excellent. Notably, 24 items with a DI of 0.19 or less raise concerns as they seem unable to effectively differentiate between higher and lower performing examinees. These items, lacking in discriminative power, should either be eliminated or undergo substantial revisions. Another set of 18 items falls within the marginal DI range of 0.20 to 0.29, indicating a need for revision to bolster their discriminatory capacity. However, there's optimism as 58 items exhibit robust discrimination abilities, with DIs ranging from 0.30 to 0.39 (good) and exceeding 0.40 (excellent). Summing up, 42 items function effectively, while 42 others show weaknesses requiring enhancement. This analysis underscores the imperative of refining over 40% of the items to ensure optimal differentiation among examinee abilities. This differentiation is pivotal for the test's validity, especially in determining pass/fail criteria.

Table 3: Item Discrimination Power

Range	**Items N=100**	**Status**	**Recommended Actions**
≤ 0.19	24	Poor	Discard/revise
0.20-0.29	18	Marginal	Revise
0.30-0.39	24	Good	Store
≥ 0.40	34	Excellent	Store

Distractor Efficiency

Table 4 provides a comprehensive overview of the quality and functionality of distractors associated with the 100 multiple choice items featured in the examination. Among the 300 total distractors examined, a noteworthy observation is that almost half (47%) underperformed as non-functional distractors (NFDs), lacking genuine plausibility as incorrect options. An encouraging finding emerged as over 50% of the items showcased perfectly functional distractors, achieving the pinnacle of distractor efficiency at 100%. However, the remaining items faced challenges, with a majority featuring one to two NFDs each, resulting in efficiency levels of 66.7% or 33.3%. Of particular concern were five items that contained three NFDs each, rendering them completely devoid of distractor functionality and yielding a disconcerting 0% efficiency.

In total, 42 items displayed suboptimal distractors, ranging from one to two per item, warranting attention for improvement. Specifically, the five items with three entirely non-functional distractors should be excluded from further consideration. Recognizing the critical role of distractor quality in accurately discerning examinee ability levels, this analysis underscores the imperative need for enhancements in nearly half of the distractors and over 40% of the items. Such improvements are essential to align with the assumptions of optimal distractor functioning outlined in Classical Test Theory.

Table 4: Distractor Analysis

Number of Items	100
Number of Distractors	300
Functional Distractors	159 (53%)
Non-functional Distractors	141 (47%)
Items with 0 NFDs (DE=100%)	53 (53%)
Items with 1 NFD (DE=66.7%)	30 (30%)
Items with 2 NFDs (DE=33.3%)	12 (12%)
Items with 3 NFDs (DE=0%)	5 (5%)
Overall mean DE (mean ±SD)	76.98±29.1

Reliability

The data presented in Table 5 provides valuable insights into the reliability of the individual test items, as assessed using the Kuder-Richardson 20 formula. The table presents a breakdown of the number and percentage of items falling within different reliability ranges. It is worth noting that out of all the test items, only 32 managed to achieve the minimum widely accepted threshold of 0.70 reliability or higher. Among these, 8 items showed excellent reliability, with a score of 0.90 or above. Additionally, 9 items demonstrated a very good level of reliability within the range of 0.80 to 0.89. Furthermore, 15 items were classified as good, falling between the reliability ranges of 0.70 to 0.79. However, it is concerning that almost half of the items (47%) only showed marginal reliability, with scores ranging between 0.50 and 0.70. More importantly, 21 items displayed poor reliability, scoring below 0.50. This raises concerns about the dependability and consistency of the measurements provided by these questions. Taken as a whole, a significant majority (68%) of the items failed to meet the 0.70 benchmark, which questions the reliability of the test. In order to ensure the psychometric soundness of the exam, it is crucial to improve over two-thirds of the test items

Table 5: Reliability of the test items

Reliability	**F**	**%**	**Status**
≥0.90	8	8	Excellent
0.80-0.89	9	9	Very good
0.70-0.79	15	15	Good
0.50-0.70	47	47	Marginal
<0.50	21	21	Poor

DISCUSSION

This study utilized classical test theory as a guiding framework to assess key characteristics of items featured in the biology teacher licensure test in Ethiopia. The obtained results provided valuable insights into the test's quality, indicating areas that require improvement.

Concerning difficulty, the average p-value of 0.53 suggests that the test achieved an overall satisfactory level of difficulty. This falls within the recommended range for tests exhibiting an expected level of difficulty ($p = 0.40$ to 0.60) (Hotiu, 2006). Notably, Mehta and Mokhasi (2014) deemed a mean p-value of 0.63 acceptable, a sentiment supported by findings of Kumar and colleagues (Kumar et al. 2014). Despite this, disparities from the recommended difficulty level distributions suggest that some items may not effectively measure the entire spectrum of candidate abilities. Approximately one-third of the items may require revision to align more closely with best practice standards. This underscores the significance of implementing quality measures in subsequent biology tests to rectify any imbalances in difficulty levels.

With regards to the discrimination index (DI), having a mean value of 0.29, the test items can be considered reasonably effective in distinguishing between higher and lower-performing examinees (Kumar, et al., 2014). However, a worrisome aspect emerges, as over 40% of items exhibited poor or marginal discrimination abilities. Given the critical role of these tests in determining licensure outcomes, addressing this weakness through item revision, particularly after a pilot study, is crucial for enhancing the test's validity. Constructing plausible distractors and minimizing non-functional distractors during the design and piloting phases of multiple-choice questions is emphasized as essential (Rodriguez, 2005).

The mean Difficulty Index (DE) in this study, at 76.98, significantly exceeds the mean DE of 47.78 reported in similar studies (Poulomi & Saibendu, 2015). Notably, only 17% of items featured two or three non-functional distractors (NFDs). Aligning with assertion of Hingorjo and Laleel (2012), items with three NFDs are considered easier, correlating with a higher mean p-value. Distractor analysis revealed that over half of the items had fully functional distractors, a positive aspect. However, almost half of all distractors and 40% of items contained one or more non-functional options, indicating areas that require improvement. Ensuring the plausibility of all distractors is crucial for maintaining the assumptions of optimal item functioning, and targeted revisions addressing non-functional distractors could enhance the performance of weakly performing items.

Moving to reliability, the optimal level is reported to be 0.70, with Feldt (2006) stating that reliabilities as low as 0.50 are acceptable for shorter tests, while tests with over 50 items should aim for reliabilities of 0.80 or higher. The reliability analysis in this study is of concern, as over two-thirds of items fall below the minimum threshold of 0.70. This inadequacy in question consistency, a key element in measurement precision, necessitates extensive revisions to elevate reliability to a psychometrically sound level. The understanding that low test reliability stems from chance differences influenced by various factors underscores the importance of addressing issues such as variations in examinee responses, psychological conditions, and poorly written or confusing items (Feldt, 2006).

The study's findings emphasized the crucial importance of integrating quality measures into the test design phase. Doing so helps identify and address potential quality issues before the tests are put into actual use. It's advisable to implement these procedures not only for this particular test but also for other assessments. Despite the valuable contributions made by this study to the existing body of knowledge, it is imperative to acknowledge certain limitations inherent in the analysis, particularly the exclusive reliance on quantitative data without the integration of qualitative insights. This narrow approach may have restricted the depth of understanding and failed to capture nuanced aspects crucial to a comprehensive evaluation.

Moreover, it is essential to recognize that the findings of the study may possess limited generalizability beyond the specific test and contextual parameters under investigation. This constraint highlights the necessity for future research endeavors to adopt a more expansive approach, employing mixed methods that encompass a diverse range of teacher licensing exams. Such a methodology promises to yield more robust and holistic evidence, enhancing the applicability and relevance of the study's findings across various educational contexts.

CONCLUSION AND RECOMMENDATIONS

Employing classical test theory as a guiding framework to assess essential characteristics of the biology teacher licensure test in Ethiopia, results of the current study offered valuable insights that can contribute to fortifying the licensure system. In general, the findings highlight certain strengths of the test, such as achieving a satisfactory level of difficulty and maintaining decent distractor quality on average. However, adherence to CTT guidelines reveals the necessity for significant improvements. Specifically, more than 40% of items exhibit weaknesses in discrimination ability, reliability analysis indicate that a majority of items fall short of accepted standards, and nearly half of the distractors require enhancement. Overall, while the present test displays acceptable attributes in certain aspects, the study results strongly imply the need for substantial enhancements in item design.

When a novice teacher license hinges solely on the outcome of a single test, devoid of any triangulated evaluation involving teacher assessments in the country, it becomes imperative to craft high-quality test items and administer them through suitable mechanisms. The key recommendations arising from this research advocate for the incorporation of quality assurance measures item analysis during test development to identify issues prior to operational use through pilot testing.

Additionally, continuous rigorous evaluation and refinement are advised to uphold defensibility over time, aligning with best practice standards. There is also a suggestion to potentially expand investigations to encompass various teacher certification exams and contexts.

Exploring alternative test construction methods, such as enhancing reliability through test lengthening if necessary, is considered. Given the dynamic nature of assessment and the crucial importance of teacher qualifications, it is implied that continuous efforts in quality improvement are essential. Adhering to these evidence-based recommendations is anticipated to fortify the defensibility and utility of licensure testing as a requirement for teacher certification.

Abbreviations

KR	Kuder Rechardson
MCQ	Multiple Choice Questions
NFD	Non-functional Distractors
DI	Difficulty Index
Rpbis	Coefficient of point biserial correlation
CTT	Classical Test Theory
DE	Distractor Efficiency
SD	Standard Deviation

ACKNOWLEDGEMENTS

The authors would like to express their sincere gratitude to the management of the Directorate for Teachers and Leaders Licensing of the Ethiopian Ministry of Education for their invaluable cooperation in providing the examination data necessary for this study. We are also thankful to the Research Ethics Committee of Dilla University for granting ethical approval and overseeing the ethical considerations associated with this research project.

REFERENCES

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics*, 25(1), 95- 135. https://doi.org/ 10.1086/508733

Anselmi, P., Colledani, D., & Robusto, E. (2019). A comparison of classical and modern measures of internal consistency. *Frontiers in Psychology, 10* https://doi.org/ 10.3389/fpsyg.2019.02714

Braun, H. J. (2005). Using student progress to evaluate teachers: A primer on value-added models. Princeton, NJ: Educational Testing Service. Available online at: http://www.ets.org/research/pic.

Brown, F. G. (2010). A short history of classical test theory. In Classical test theory. SAGE Publications.

Eapen, D. (2017). A guide to practical human reliability assessment. CRC press.

Egun, N. K. (2016). *Teacher Qualification and Students Performance in Biology: A Study Schools in Ethiope East Local Government Area of Delta State.* Delta State University, Faculty of Education. Abraka: Delta State University.

Feldt, L. S. (2006). Reliability coefficients, Kuder-Richardson. Encyclopedia of Statistical Sciences. https://doi.org/10.1002/0471667196.ess1370.pub2

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items*. Routledge, New york. Available at https://doi.org/10.4324/9780203825945

Haladyna, T. M., & Rodriguez, M. C. (2021). Using full-information item analysis to improve item quality. *Educational Assessment,* 26(3), 198-211. https://doi.org/ 10.1080/ 10627197.2021.1946390

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational measurement: Issues and practice*, 12(3), 38-47.

Harris, D. N., & Sass, T. R. (2007). *Teacher training, teacher quality, and student achievement. Working paper 3*. https://files.eric.ed.gov/fulltext/ED509656.pdf

Hingorjo, M. R, & Laleel F. (2012). Analysis of One -Best MCQs: the Difficulty Index, Discrimination Index and Distractor Efficiency. *J Pak Med Assoc*, 62, 142-147. https://educ5let2012.blogspot.com/2012/09/item-analysis-and-item-discrimination.html

Hotiu, A. (2006). The relationship between item difficulty and discrimination indices in multiple-choice tests in a Physical science course, MSc thesis, Boca Raton, Florida: Florida Atlantic University. Available at: http://www.physics.fau.edu/ research/education/A.Hotiu_thesis.pdf

Instructional Assessment Resources [IAR]. (2011). Item Analysis, University of Texas at Austin, Instructional Assessment Resources (IAR) Web site: http://www.utexas.edu/ academic/ctl/assessment/iar/students/report/ itemanalysis.php, Retrieved November 9, 2013

Kirwan, B. (2017). Human reliability assessment in risk assessment. A Guide to Practical Human Reliability Assessment, 17-38. https://doi.org/10.1201/9781315136349-4

Kumar, P., Sharma, R., Rana, M., & Gajjar, S. (2014). Item and test analysis to identify quality multiple choice questions (MCQS) from an assessment of medical students of Ahmedabad, Gujarat. *Indian Journal of Community Medicine*, 39(1), 17. https://doi.org/10.4103/0970-0218.126347

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Addison-Wesley, Reading.

Mehta G, Mokhasi V (2014). Item analysis of multiple choice questions- an assessment of the assessment tool. *International Journal of Health Sciences Research4*(7):197-202. https://www.ijhsr.org/IJHSR_Vol.4_Issue.7_July2014/30.pdf

Mensah, R. O., Acquah, A., Frimpong, A., & Babah, P. A. (2020). Towards improving the quality of basic education in Ghana. Teacher licensure and matters arising: Challenges and the way forward. *Journal of Education & Social Policy*, 7(3). https://doi.org/10.30845/jesp. v7n3p11

Ministry of Education (1994). Education and Training Policy, Addis Ababa.

Ministry of Education (2010).Education Sector Development Programme –IV, Addis Ababa.

Ministry of Education (2015).Education Sector Development Programme – V, Addis Ababa.

Mitchell, R. & Barth, P. (1999). Not Good Enough: A Content Analysis of Teacher Licensing Examinations. How Teacher Licensing Tests Fall Short. *Thinking K-16*, *3*(1), 3-21. https://eric.ed.gov/?id=ED457261

Moore, S., Fang, E., Nguyen, H. A., & Stamper, J. (2023). Crowd sourcing the evaluation of multiple-choice questions using item-writing flaws and bloom's taxonomy. *Proceedings of the Tenth ACM Conference on Learning @ Scale*. https://doi.org/10.1145/3573051.3593396

Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. Review of educational research, 62(3), 229-258. https://doi.org/10.3102/00346543062003229

National Research Council. Committee on Assessment and Teacher Quality. (2001). *Testing teacher candidates: The role of licensure tests in improving teacher quality*. National Research Council. https://nap.nationalacademies.org/download/10090#

Poulomi M, Saibendu K. L. (2015). Analysis of Multiple Choice Questions (MCQs): Item and Test Statistics from an assessment in a medical college of Kolkata, West Bengal. IOSR Journal of Dental and Medical Sciences. 14(12). 47-52. www.iosrjournals.org

Raîche, G. (2005). Critical evaluation of classical test theory procedures used in the study of equivalent test forms. Applied Psychological Measurement, 29(5), 393-409. https://doi.org/10.1177/0146621604271596

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review*, 94(2), 247-252. https://doi.org/10.1257 /0002828041302244

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice*, 24(2), 3-13. https://doi.org/10.1111/j.1745-3992.2005.00006.x

Rogers, A. P., Reagan, E. M., & Ward, C. (2020). Preservice teacher performance assessment and novice teacher assessment literacy. *Teaching Education*, 33(2), 175-193. https://doi.org/10.1080/10476210.2020.1840544

Suciati, Munadi, S., Sugiman, & Febriyanti, W. D. R. (2020). Design and validation of mathematical literacy instruments for assessment for learning in Indonesia, *European Journal of Educational Research*, 9(2), 865-875. https://doi.org/10.12973/eu-jer.9.2.865

Talal, A, Peter T, Per Kind (2015). The Tools of Teacher Evaluation: What Should Be Used in Teacher Evaluation from the Teachers' Perspective. Durham University, Conference Paper. UK,. International Business & Education Conferences – June 7-11 London, UK

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha, *International journal of medical education*, 2, 53-55. https://doi.org/10.5116%2Fijme.4dfb.8dfd