DECODING TEST ITEMS: THE ART AND SCIENCE OF ITEM ANALYSIS

Abstract

Item analysis plays a crucial role in evaluating the quality and effectiveness of test items. Through the examination of item difficulty, discrimination, and distractor analysis, educators can make informed decisions about the reliability and validity of their assessments. By continuously reviewing and refining their test items, educators can promote fair and accurate evaluation of student knowledge and skills.

The analysis of the items is aimed at determining how well a set of items functions both in terms of their psychometric properties and their content and educational relevancy. Item analysis is concerned with ensuring the relevance and effectiveness of test items.

Authors

Kiran Kumar Ganji
Institute of Health Professions Education,
Sri Balaji Vidyapeeth,
Pondicherry, Tamilnadu, India
kiranperio@gmail.com

Prof. N. Ananthakrishnan
Emeritus Professor, Surgery and HPE,
Institute of Health Professions Education
Sri Balaji Vidyapeeth
Pondicherry, Tamilnadu, India
n.ananthk@gmail.com

I. IMPORTANCE & RELEVANCE

One tool that teachers frequently use to gauge their students' learning outcomes is a test. According to Brown, it is described as “a method of measuring a person’s ability, knowledge, or performance in a given domain”.[1] "A test is an instrument - a collection of techniques, procedures, or items that requires performance on the part of the test taker," Brown continued. A test is "an effort to determine the student's status in terms of their knowledge, skills, and attitudes," according to Popham.[2] Tests offer useful data on many facets of the teaching-learning process, which can be utilized to assess the teaching-learning program itself, according to Bachman & Palmer.[3] They go on to explain that tests can offer the following: a diagnosis of strengths and weaknesses that helps determine whether a class or individual students are ready to move on to the next unit of instruction; a method of assigning grades based on learners' achievement; proof of the outcome of learning and teaching that can serve as feedback of the efficacy of the teaching program itself; information used to decide what kinds of learning materials and activities that should be given to students; and a way of clarifying the instructional objectives, instructional materials, and activities based on the students' need for learning. Teachers must ensure that the tests they create are of high quality, given the test's pivotal position in the educational process. According to Weir, test developers and instructors must make sure that test scores accurately represent an examinee's aptitude in a particular subject.[4] While creating test items appears to be very simple, designing them right takes a great deal of effort and time. Since each assessment tool has its advantages, no single instrument is ideal or able to fully capture all facets of students' performance and competency as well as shortcomings.[5] Item analysis is one technique to make sure an exam is of high quality. A series of steps called item analysis are used to assess the caliber of the test items.[6] "Item analysis is typically done for the purpose of selecting which items will remain on future revised and improved versions of the test," according to Brown & Hudson.[7]

Item analysis was defined as a method based on certain procedures and steps to determine which test items are efficient and of high quality to be utilized as an evaluation tool by Brown & Hudson [7] and Musial.[6] A post-examination assessment called item analysis can reveal details regarding the caliber of the tests. A statistical analysis of a test taker's responses is called item analysis. Gathering and compiling student answers can yield quantifiable, objective data that is helpful in determining the caliber of the test items and boosting the effectiveness of the evaluation.[8, 9] Item analysis also "examines how well individual items perform when compared to other test items or to some external criterion".[10]

Item analyses of large-sample test data, whether constructed by a test author or purchased from an external publisher, have automatically transformed into a necessary step resulting in the production of diagnostic performance reports. Test authors are trained in the interpretation and understanding of test performance reports. In contrast, individuals who acquire tests, such as instructors, administrators, or educators, interested in interpreting the statistics and indices provided within the report, normally receive little or no instruction with regard to the report's interpretation Test statistics frequently are reported but usually are misunderstood, thus making them a complete or erroneous misinterpretation. So, it is vital to explicitly guide those interested in learning how to assess the performance of test items, using item analysis as a mechanism.

Overall, item analysis plays a crucial role in evaluating the quality and effectiveness of test items. Through the examination of item difficulty, discrimination, and distractor analysis, educators can make informed decisions about the reliability and validity of their assessments. By continuously reviewing and refining their test items, educators can promote fair and accurate evaluation of student knowledge and skills.

1. Purpose and Importance of Item Analysis

Decoding test items, the basic building blocks of educational and psychological assessment, is a process that can be thought of as both an art and a science. The goal of this process is to evaluate the effectiveness of test items and their relevance to the assessment objective or goal. This monograph highlights the relevance of the analysis in quantitative terms, as well as its importance in qualitative terms despite the lack of usable statistics. This monograph also presents a review of various statistical item analysis indices from the most basic to the more sophisticated indices, which attempt to model item response functions across a continuum of examinee ability, or trait. Some indices can easily be computed even without any expert computational package on educational assessment.

A test item can be defined as any stimulus (question, opinion, situation, illustration) employed to elicit a response from an individual with the goal of drawing a conclusion about only one or a limited number of characteristics of that individual. For research purposes, an assessment may be defined as an organized collection of work that represents the evaluation of an individual's performance or progress in a given area. This can be done independently or through the use of assessment centres, as in the case of psychological testing. An educational or psychological test is a form of assessment, but not all assessments are tests. Tests are formal procedures that adopt a set of instruments, predetermined methodologies, norms and unequal attention to test takers in order to make decisions about examining individuals.

The item analysis consists of a three-level performance assessment of test items. The first level of analysis includes overall test statistics and the typical reporting of a large-sample test. This often includes the number of items and the number of students tested, the number omitted, the mean raw score, the standard deviation of the raw scores, the reliability of the raw scores, the percentage of students scoring zero, and the percentage scoring full credit. In addition to these, test statistics based on the dichotomous response were reported. The second level of analysis, a more detailed reporting of the test performance, is what is typically thought of as item analysis. It is the reporting of the number of students who answered each item correctly or incorrectly and the percentage of students who accomplished this. Another aspect of this analysis is the item difficulty, which expresses how easy or difficult each item was relative to the other items in the test.

Item difficulty is a crucial measure in item analysis as it allows educators to evaluate the discriminative power of each item in the test. By examining the item difficulty, educators can gain insights into the effectiveness of their test questions and identify areas that may require further attention in instruction. The item difficulty is calculated by dividing the number of students who answered the item correctly by the total number of students who attempted the item.

Furthermore, item discrimination is another key component of item analysis. Item discrimination measures the ability of an item to differentiate between high-performing and low-performing students. It provides valuable information on how well an item discriminates between students who have mastered the content and those who have not. High item discrimination indicates that the item effectively distinguishes between students with different levels of knowledge or skill. To determine item discrimination, various statistical methods can be employed, such as the point-biserial correlation or the phi coefficient. These measures assess the relationship between an item response and the total test score. A positive correlation indicates that students who answered the item correctly performed better on the overall test, while a negative correlation suggests that students who answered incorrectly performed better. In addition to item difficulty and discrimination, the distractor analysis is another important aspect of item analysis. Distractors are the incorrect options provided alongside the correct answer in multiple-choice questions. Through distractor analysis, educators can evaluate the effectiveness of these incorrect options. The analysis involves determining the percentage of students who selected each distractor, as well as the percentage of students who selected no response or omitted the item entirely.

By examining the performance of each distractor, educators can identify common misconceptions or misunderstandings among students. This information can guide the revision of test items and help improve the overall quality of assessments. Additionally, the analysis of omitted responses provides insights into the factors that may influence student decision-making or test-taking strategies.

2. Historical Development

The theory of item analysis has matured from the older "tricks of the trade" to become a science built on accepted principles with its own professional literature. A theory is more than the formulae, tables, and charts used; it includes an understanding of the phenomena to which those instruments apply and the tacit rules employed in their use. Earlier in the history of testing, the available tools and techniques for the manipulation of items and tests were more or less serendipitous.[11] There were some expert test item writers and testers who acted as the "wizards" successfully manipulating items and tests without the underlying rationale of their performances being understood by others. Experts paid close attention to knob turning and the rules embodied in the knob-turning. For example, Ebel presented a set of item writing and selection strengths that were based on empirical observations and correlations from which they might be eventually proved by science. On the other hand, there were 1920s textbooks on the art of item writing and promising to turn testers into master manipulators of items and tests by watching for the implemented sample formulations like item writing tricks.[11]

Except in a few controlled testing situations, the naïve testers were doomed. But the development of modern item analysis tools freed them from testing wizardry and made high stakes testing with the use of item analysis tools accessible to the masses. In a parallel manner the development of the item analysis software for the microcomputers has made item analysis technology accessible to the users who have little or no intention to understand the rationale underlying the use of the tools. Like other technologies, item analysis software is also susceptible for misuse. A lack of knowledge about how the fundamental assumptions in

item analysis are violated may result in serious mistakes and conclusions regarding the characteristics of the tests and items.

II. ITEM ANALYSIS PROCESS

The item analysis process is comprehensive and elaborate by necessity, encompassing multiple steps. As an initial step, data on the test items must be compiled and recorded in a computer file for effective calculations. Methods and software programs for scoring or analyzing items can vary significantly. Lotus 1-2-3, Microsoft Excel, SPSS, SAS, and the statistical software programs provided by software companies are now commonly used for data analysis. There are probably other "home-grown" computer programs available local to institutions that would do a similar job. As with examining the quality of individual items, it is important to recognize the differences between item scoring and item analysis.[12]

Item scoring refers to automation of continually presenting questions, collecting responses, and reporting scores. Item analysis means interpreting test data in terms of the quality of items or test. A thorough item analysis involves computation of many item statistics. The types of calculations performed depend on the test format and purpose for the item analysis. There is, however, a core group of statistics that are reported for almost all tests. These include descriptive statistics such as item total and percent correct, item facility and discriminability indices, slopes of the item characteristic curves, Rasch item and test person parameters, possibly some chi-square goodness of fit statistics, etc. Regardless of its sophistication, analysis output must be in an interpretable form involving summary statistics to avoid "getting lost" in numbers.[11] Generally, frequency distributions, means, confidence intervals, and standard deviations are computed for continuous data. These data can be quickly summarized in tabular form and/or visually presented through histograms or boxplots for easier interpretation. A basic item analysis comprises several tasks: computation of basic item statistics, the interpretation of statistics in terms of the quality of items, the identification of serious problematic items, and the recommendation for item revision.

1. Data Collection

Focuses on item analysis, primarily associated with multiple-choice items. Includes an overview of strategies used for gathering needed data and activity. To perform an item analysis of a set of test items with an emphasis on multiple-choice items. It provides a thorough review of several approaches to data collection and a discussion of the strengths and weaknesses of each procedure for item analysis of the logic of science based (LS) test. Item analysis is a method of evaluating the performance of test items to improve the validity, reliability, and fairness of a test. It involves gathering statistically observable data about test performance, along with specific rules for interpreting the data to compute various item characteristics.[12] The rationale for examining test items is based on general principles from educational measurement theory. In classical test theory (CTT), each item is evaluated based on its ability to discriminate between students who know the subject and those who do not.[11]

2. Calculation of Item Statistics

Covering a crucial aspect of decoding test items, the objective is to provide insights into the quantitative measures utilized in item analysis to assess the characteristics and performance of test items. Item statistics, which play a vital role in the evaluation process, are calculated to determine the degree of item difficulty and to identify items that do not conform to acceptable levels of difficulty. In addition, the Discrimination Index and the Point Biserial Correlation Coefficient are computed to identify items that do not conform to acceptable levels of discriminating power.

The classical method of Item Analysis comprising three basic steps—calculation of Item Statistics, elective appraisal of items according to the item statistics, and presentation of results—has been employed in this study. Item statistics including the Facility Index, Discrimination Index, and Point Biserial Correlation Coefficient measures.

3. Key concepts in Item Analysis

- **Reliability and Validity:** Test items, whether in paper-and-pencil, computer-based, or performance formats, must be reliable and valid measurement instruments. The accuracy of the score assigned to these test items is paramount, as is the presence of these test items on the assessment. Reliability looks into the random error in measurement, while validity assesses how accurately a test measures what it is supposed to measure.[13] Content and construct Validity (a technique analyzes what it wants to assess), reliability (the degree to which a score accurately represents an individual's abilities), and objectivity (an evaluation with a single correct answer) are all necessary components of any assessment. These characteristics can be recalled using higher-order thinking skills and problem solving. Every assessment technique has both advantages and limitations.[14] An assessment tool's item analysis provides information on an item's validity and reliability.

- **Reliability:** Test score Reliability is the likelihood that scores will remain consistent over time if the same exam is administered to the same students several times. Cronbach's Alpha is a measure of internal consistency that provides dependability data for items evaluated dichotomously (correct/incorrect), such as multiple-choice questions. A test with a Chronbach's Alpha score of.80 or more has lesser measurement error and is therefore regarded as having extremely strong dependability. A value less than .50 is considered to have low reliability. Item The dependability of your test reflects how well it assesses learning on a single topic. Internal consistency measures indicate how consistently and collectively the test's questions target a common topic or construct. Reliability is crucial. The reliability coefficients range between 0.00 and 1.00. Ideally, score reliability should be greater than 0.80. Coefficients in the 0.80-0.90 range appear to be ideal for course and licensure exams.

The test's correlation with itself is used to interpret reliability. The percentage of a test score attributed to mistake will go down as reliability estimates rise. Understanding how connected the items are to one another and if they measure a single latent feature or concept is necessary for wise alpha interpretation. Exam or examination with varying content materials, such as integrated courses. For instance, the musculoskeletal system course includes various topics from fundamental medical and clinical sciences that have distinct contents, even if anatomy

dominates the course. As a result, interpreting a course test like this requires careful consideration beyond the alpha figure. According to reports, a short test (less than 50 items) with a KR20 of 0.7 is acceptable, while an extended test (more than 50 items) with a KR20 of 0.8 is acceptable.[15] Furthermore, research has shown that a multidimensional exam's alpha value is not lower than a unidimensional exam's.[16] A low alpha value may result from diverse constructs, fewer items, or less interrelatedness between items.[17] Exam dependability may be indicated by a high alpha value, and certain items are non-functional because they test the same material repeatedly or in a different format.[17, 18] Furthermore, a high value denotes highly connected elements, suggesting a constrained coverage of the content materials.[17] Test reliability can be raised by including more items with a tolerable difficulty index, strong discrimination power, and distractor efficiency.[17, 19, 20] Furthermore, removing items that are flawed or have a p-value that is too high or too low might raise Cronbach's alpha. Exam items that have weak correlation or are unrelated should be changed or removed.

Classification of KR20 Value and Its Interpretation

Table 1

KR20 value	Interpretation of Cronhbach's alpha (KR20)	Author
≥0.80	Exemplary	Robinson, Shaver et al.[21]
0.70–0.79	Extensive	
0.60–0.69	Moderate	
<0.60	Minimal	
<0.70	Unacceptable	Cicchetti[22]
0.70–0.80	Fair	
0.80–0.90	Good	
< 0.90	Excellent	
>0.90	Needed for very high stakes tests (e.g., licensure, certification exams)	Axelson and Kreiter[23]
0.80–0.89	Acceptable for moderate stakes tests (e.g., end-of-year summative exams in medical school, end-of-course exams)	
0.70–0.79	Acceptable for lower stakes assessments (e.g., formative or summative classroom-type assessments created and administered by local faculty	
<0.70	Useful as one component of an overall composite score.	
>0.90	Excellent reliability	Obon and Rey([24]
0.80–0.90	Very good for a classroom test	
0.70–0.80	good for a classroom test	
0.60–0.70	Somewhat low (The test needs to be supplemented by other measure)	
0.50–0.60	Suggests need for revision of test (unless it is quite short, ten or fewer Items).	
<0.50	Questionable reliability.	

KR20 value	Interpretation of Cronhbach's alpha (KR20)	Author
> 0.7	Excellent	Hassan and Hod[25]
0.6–0.7	Acceptable	
-0.5-0.6	Poor	
< 0.5	Unacceptable	
< 0.30	Unreliable	

Reliability and validity are crucial for identifying the outcomes that meet the standards and evaluating bias. Reliability reveals the level to which assessments were consistent, whereas validity investigates assessment correctness.[26] Internal consistency, stability, equivalence, and precision are all terms for reliability. The standard error of measurement and the standard deviation of the examinee's assessment are both elements that influence reliability. Internal consistency is estimated using the item's average correlation for a test and the extent to which the MCQs may assess the same knowledge domain aspects. Internal consistency is frequently determined by calculating the reliability coefficient. A reliability coefficient calculates the concordance between examinees' observed and true scores, as well as the relationships between scores obtained from two parallel exams. This estimate explains why an individual's scores are likely to fluctuate when retested with the same or comparable test, assuming no change in knowledge or perception.[27-29] Increasing the number of items in a test can improve reliability, but it is expensive, time-consuming, and necessitates an average correlation effort.

The reliability score is influenced by a number of factors, some of which are under your control and others not.

Table 2

Factor	Why it's important
Duration of the examination	The inclusion of more items enhances reliability.
Percentage of pupils answering each item correctly and incorrectly	Assists in assessing the item reliability.
Difficulty level of an item	Items that are either very easy or very tough do not effectively differentiate across individuals and will decrease the reliability estimate
Homogeneity	Including more items on a topic increases reliability. This can be difficult when a test covers many areas. Ask questions that are varied enough to survey the topics but similar enough to represent a certain theme.
Number of individuals taking the test	The reliability of test results increases as the number of students being assessed using the same set of items increases.
Variables that impact an individual test taker on a specific day	Factors such as preparedness, distraction, physical fitness, and exam anxiety can impact students' capacity to select the appropriate option.

There are five commonly used item statistics—item difficulty, item discrimination, item pseudo-guessing, point-biserial correlation, item reliability, and item validity. These item statistics provide important information about the performance and quality of an item in a assessment or test. Understanding these statistics can help educators and researchers to make informed decisions about the validity and reliability of their assessments. Item difficulty refers to the proportion of test-takers who answered the item correctly. It gives an indication of how challenging the item is and how well test-takers are able to understand and respond to it. Item discrimination, on the other hand, measures how well an item distinguishes between high-performing and low-performing test-takers. It helps to identify items that are effective in differentiating between individuals with different levels of ability or knowledge. Item pseudo-guessing is a statistic used in multiple-choice items to estimate the probability of getting the answer correct by guessing alone. This helps to determine if guessing is a significant factor that influences the test scores and affects the validity of the assessment results. Point-biserial correlation is a statistical measure that assesses the relationship between an item score and the total test score. It shows how well an item correlates with the overall performance on the test, indicating the contribution of the item to the test's reliability. High point-biserial correlation indicates that the item is measuring the same construct as the overall test and contributes to the reliability. Item reliability measures the consistency or stability of an item over repeated administrations. It is crucial for ensuring that the assessment produces reliable and consistent results. Reliability can be assessed using various methods, such as internal consistency and test-retest reliability. Item validity refers to the extent to which an item measures what it is intended to measure. It assesses whether the item accurately captures the construct or concept it is designed to represent. Validity is fundamental for ensuring that the assessment results are meaningful and accurate. These commonly used item statistics provide valuable insights into the properties of assessment items. By examining these statistics, educators and researchers can make informed decisions about item selection, modification, and overall test improvement.

❖ Validity Evidence

Validity in assessment is not a unitary concept—there are multiple types of validity evidence that are necessary to determine whether each test item is sufficiently relevant, appropriate, and adequate for the inferences intended. The need for specific types of validity evidence can be traced to the inferences drawn from assessment outcomes. These inferences can be grouped along three dimensions: 1) inference from point estimates derived from the entire test to the underlying knowledge, skill or ability being measured; 2) inference from overall test results to knowledge, skill or ability in specific content domains; and 3) inference from overall test results to predicted level of performance in a specific setting. Each of these inferences is commonly attended to in the assessment literature.[30]

An important distinction that drives decision making regarding test items is that between evidentiary basis and strength of evidence. Although there are many different ways to establish each type of validity evidence, they can be grouped into roughly two categories: 1) protocols or procedures, which detail the collection or evaluation of data; and 2) resulting validity evidence, which refers to the statistical or descriptive indices summarizing the analysis. Each category has in turn been divided into broad types of criteria used as evidence for validity.[31]

❖ Item Difficulty

Overall, the items in a test should be of appropriate difficulty for the intended target population. To achieve this, there is a need to examine the difficulty of each item using the statistic of item difficulty. It is useful to know which items are more difficult than or easier than the intended difficulty. In general, for a well-constructed test on the intended level of difficulty, approximately half of the items should be of moderate difficulty. Difficulty levels can also be determined by other characteristics, such as item types, subcategories, and item writers. However, the degree of difficulty is not revealed by any of the characteristics to make sure that test developers can exclude certain items from consideration. Based on the expected overall item difficulty of a test, an optimal number of items can be pre-calculated to be more difficult than or easier than this level of difficulty. With the item statistics after the test administration, the items on both sides of the intended item difficulty can be identified.

❖ Item Discrimination

In general, tests are constructed targeting a certain population, such as students who have taken the particular course, and items are expected to be answered correctly by some of these individuals. To determine if items are behaving differently from the student population in general, there is a need to examine item discrimination. Overall, the items in a test should have a certain level of discrimination from the non-target population. To achieve this goal, the statistics of item discrimination can be computed. It is useful to know both the items that have higher item discrimination than the intended level and the items that have lower item discrimination than this level. In general, for a well-constructed test targeting the intended population, approximately half of the items should have moderate item discrimination. Item discrimination reflects how well each item distinguishes between the respondents who are more likely to pass or fail the test. There is a need for two types of interpretation for item discrimination: dichotomous scoring and positive scoring. Understanding item discrimination with regard to point-biserial correlation can also provide better insights about item discrimination.

4. Identification of Problematic Items

With respect to criterion referenced tests while looking at the entire test it is difficult to know which items are problematic and need further investigation. To enable item analysis the problems with respect to questionable or defective items are either put on an item by item basis or based on a set of criteria which are procedures and methods of recognizing items within the assessment that are to be looked into more closely for further evaluation or modification.

In a subjective test even a great test with uniformly difficult items cannot be expected to yield a normal probability curve spread of marks. Even in a multiple choice test a question need not invariably have only one of the four options correct. The option should pertain to a particular context which is precise at least from a latter view.[32] At least on examination of a large number of questions it would be normal to expect some common defects.[33] The problematic items may be grouped as faulty or ambiguous questions, faulty or ambiguous options, too simple or too difficult questions, technical problems, silly mistakes in scoring key and items not conforming to specification.

5. Item Revision Strategies

Refinements and Enhancements. Item revising can involve editing test items or item writing, but the focus of this discussion is on item editing. Item revision is the process of innovating an existing item under analysis, and often follows an item analysis. This could involve improving a poorly performing item based on data, modifying an item to improve clarity, grammaticality, independence, and bias, or such changes as those made merely to maintain the yardstick of the item pool after repeated use.[34]) When done based on conclusions drawn from an item analysis, item revision is a form of item enhancement and item repair.

Multiple Approaches. There is a belief that item analysis is only of limited help in developing effective tests because items can be misused or designed in malfunctioning ways as well as used or constructed properly without falling into malfunctioning categories. Such invalid items remain valid despite their flaws, while other valid items may non-frustrate the instructional goals of a particular assessment. In the long run, item faults may frustrate the purpose of assessment. However, such items may nevertheless not malfunction in a design sense, while unfortunate designs may still lead to desirable outcomes. Therefore, validity should not be solely dependent on item analysis. A more appropriate approach might be to take multiple perspectives in conjunction with data analysis. From a broader viewpoint of socio-constructivism and consequential validity, items could be treated as information sources or argumentative premises with different affordances and challenges. Item validity might be operationalized as epistemic efficacy and concerned with the degree to which an item coherently behaves in terms of instructing desired performances. It is also critical to consider the implementation context (e.g., role of items in social agendas and tests) and the historic trajectory of an item pool (e.g., where its success remains unchallenged).

III. ITEM DIFFICULTY

A common way to assess the quality of multiple-choice test items against norms is to look at the degree of item difficulty. It is typically expressed in terms of a difficulty index. Some items are very easy, others very hard, and the great majority of items are neither too easy nor too hard.

Item difficulty is defined as the proportion of examinees who answered the item correctly, and it is usually indicated by the symbol p (the lower case of P). For a multiple-choice item with a single correct response, it is calculated as follows:

$$p = (\text{number answering correctly})/(\text{total number answering})$$

There are limits on the computed value of p. In the extreme case where all examinees chose the correct answer, p = 1.00, which indicates a very easy item. Conversely, in the case where no examinees chose the correct answer, p = 0.00, indicating an excessively hard item. Item difficulty p values between .30 and .70 are generally considered to be the optimal degree of difficulty (or discrimination). p values below.[20] suggest that items may be too difficult for the sample, while p values above .80 indicate that items may be too easy for the group. When items fall below the established range, they will need to be reviewed carefully in conjunction with option analysis.

The item difficulty index offers several advantages. First, it provides information that helps to frame a test in order to maximize its reliability or predictive power. Second, it can help to identify items with problems that need to be investigated further. Some items may encounter difficulty ratios that are out of the ordinary relative to the group norm and might suggest a negative impact on test performance.

1. Definition and Calculation

In classical test theory, it is assumed that test takers are drawn from a larger population, and that test items (questions) are randomly drawn from a larger pool of items, as if each behavior in that larger population is equally likely to occur. Whether intentionally or unintentionally, every test that is created designs and oversees the population from which respondents are drawn. When the test scores of the test takers are compared to those of the larger, general population for that test, it is assumed by the test maker that the test takers are similar to others within that population. It is also commonly assumed that the items in the pool of items are generally similar in their ability to discriminate between performances of test takers from the population.

Item difficulty refers to how many test takers correctly answer a test item. A test item is "easy" if a high proportion of the test takers answer it correctly; a test item is "difficult" if a low proportion of the test takers answer it correctly. Item difficulty values range from 0 (no one answers the item correctly) to 1 (everyone answers the item correctly). For an item that can be answered correctly in only one way (for example, a multiple-choice item with only one correct choice), the probability of a correct response is the proportion of the target population that is assumed to correctly answer it.

Let c represent the number of individuals from the target population whom the item correctly classifies, n represent the total number of individuals from that population, and r represent the difficulty level of the item. The item analysis may be summarized using the following:

- Item Difficulty (p). An assessment of item difficulty based on the number of correct responses (c) within the total number of responses (n). The value of c represents the number of individuals who have responded correctly to an item. The value of n represents the total number of individuals within a group whose responses for the item are being evaluated. The value of p is the proportion of individuals responding correctly to the item, calculated from the ratio of c/n. The closer the value of p is to 1, the easier the item. -Calculation of Index of Difficulty: Difficulty index or P value using formula $P = H + L/N \times 100$

  H = The count of students who answered the item correctly within the high achieving category.

  L = The count of students that responded correctly to the item within the low achieving category.

  N = The combined total of students in both groups, including those who did not answer.

2. Interpreting Difficulty Indices

Although they seem straightforward, difficulty indices need careful thought in their interpretation. The meaning of absolute values is difficult to determine because, for instance, it is uncertain what it means if an index value is below 0.10 or above 0.90. Regarding relative values, it can be certain that for a set of tests and their items all with the same method of determining difficulty, the order of difficulty indices is accurate. However, to what extent tests or items are more difficult or easier than the average test or item is impossible to determine.

Generally, on a multiple-choice test of achievement, indices above 0.90 or below 0.10 are unusual and should raise a red flag signal. With a difficulty index of greater than 0.90, it is likely that an item is too easy or has been poorly discriminated from the test as a whole. With an index of less than 0.10, it is likely that an item is too difficult, has been poorly discriminated from the test as a whole, or has been poorly constructed. Such indices merit special attention from the test constructor.

Table 3: Various indices for interpretation of difficulty index (power)

Difficulty power	**Interpretation**	**Author**
>80%	Easy	Uddin et al.(35)
30–80%	Moderate	
<30%	Difficult	
>80	Easy	Kaur, Singla et al.(36)
40–80	Moderate	
<39	Difficult	
90	Easy	Sugianto(37)
50	Moderate	
10	Difficult	
<30	Too difficult	Date, Borkar et al.(38) and Kumar, Jaipurkar et al.(39)
>70%	Too easy	
50–60%	Excellent/ideal	
30–70%	Good/acceptable/average	
> 0.76	Easy (Revise or Discard)	Obon and Rey(24)
0.26–0.75	Right difficult (Retain)	
0–0.25	Difficult (Revise or Discard)	
>70%	Easy	Bhat and Prasad(32)
30–70%	Good	
<30%	Difficult	

If a set of parallel forms of a test has been individually administered to matched groups of subjects, indices of the forms can be examined for the indices of the items on those tests. Here again pondering on the absolute values of the variance is not worthwhile. However, if the absolute values of the difficulty indices are compared first to see if they are both above 0.90 or both below 0.10 and then check to see if there is a large difference between them, that would give some impetus to the investigation. Such considerations should be further explored with the help of histograms. Similarly, for a set of items, often called "a bank," with the same parallel form of a test, if the difficulty indices show a marked discrepancy such that a set of

items is either all much more difficult or much easier than all its counterparts on the original test, it may warrant some investigation of those items.

3. Item Discrimination

Test items can either be favorable or unfavorable to students. Favorable items are those that a high proportion of students got correct and unfavorables items, on the other hand, are those that a high proportion of students got wrong. Item discrimination concerns the interpretation of test items that fall in the given categories. Items that are disproportionately easy are weak items since they cannot discriminate between high and low performers. A test item L that is easy for all students will be weak since it does not discriminate high and low performers. Test items that are disproportionately difficult can be viewed in two ways. Such items can be considered penalty items since they are favorable to those who perform poorly. Such items can also be viewed as weak since they allow high performers to attempt them unsuccessfully. A test item K that is difficult and wrong for all students cannot operate as a discriminator between individuals. Item analysis is concerned with the performance of items on a test, and as such, it draws attention to the functioning of items in a distribution of total score performance. Several measures have been proposed that assess how well items differentiate among individuals with high and low total test scores [6].

The point biserial index of discrimination is widely accepted as often used (for n = number of students, c = number of n given correct, r = mean scores of n on p, and s2 = variance of scores of n on p) to calculate item discrimination. The point biserial index is the difference between the mean scores of the two individuals with respect to a dichotomous variable j (on p i.e x = 1 or 0), multiplied by a correlation coefficient (r = point biserial index), and standardized with respect to the dependent variable. With respect to question (item) p, xk is 1 for those with correct score and 0 otherwise. Discrimination indices are interpreted as follows: $r < 0$ (negative discrimination): the item is answered correctly by a higher proportion of those with lower total test scores than those with higher total test scores, $r = 0$ (zero discrimination): the item is equally answered correctly by those with high total test scores and those with low total test scores, and $r > 0$ (positive discrimination): the item is answered correctly by higher proportion of those with higher total test scores than those with lower total test scores.

❖ Definition and Calculation

The term item discrimination often refers to a total test score divided by the number of items and can be interpreted in several ways [6]. In most contexts, discrimination values for test items are interpreted as indicators of a test item's ability to discriminate between the underlying knowledge, skills, and abilities of the test takers being assessed. The more candidates know, the higher probability that they answer the question correctly. Items that are answered correctly by a high proportion of able candidates and incorrectly by a low proportion of less-able candidates are termed "highly discriminating" items. These items reflect candidates' underlying knowledge and, thus, assess the same "thing." In contrast, items that are answered correctly by a low proportion of able candidates and incorrectly by a high proportion of less-able candidates are termed "poorly discriminating" items. These items may be flawed and would not be expected to accurately assess the underlying knowledge, skills, and abilities being measured by the test.

Calculation of Discrimination index (D) or d value using formula, d = H – L ×2/N

H = The count of students who answered the item correctly within the high achieving category.

L = The count of students that responded correctly to the item within the low achieving category.

N = The combined total of students in both groups, including those who did not answer.

There are a variety of methods used for calculating item discrimination values. The most straightforward procedures derive a discrimination value for each item by computing the correlation between candidates' scores on each item and their total test score. When using this method, it is necessary to first adjust total test scores so that they reflect only those items being used for the discrimination calculation. Total test scores that include an item in question will yield an artificially high correlation because candidates who answer the item correctly or incorrectly will also score higher or lower, respectively, on the items that were used to derive the total test score. Therefore, it is essential that the discrimination value calculation be based on test scores reflecting all items except the item being examined for discrimination.

❖ Interpreting Discrimination Indices

Discrimination indices reflect the ability of a test item to distinguish those who do well on the overall assessment from those who do poorly, and some calculation processes are prerequisite. In general, discrimination indices can be computed for categories of items and an entire test, as well as for individual items.[40] Item discrimination is a characteristic of the item, not the test, and hence is expressed as a statistic appropriate for indices used to report reliability or difficulty. Two effective means of item discrimination are the point-biserial (pb) coefficients and the biserial (pbis) coefficients. In the case of tests with balanced right and wrong answers, to ensure continuity with the point-biserial correlation, the average trues of the IRT theta estimate will be close to zero yielding an average pb whose original estimate is also on the interval. Barely acceptable pbs of a weighted-item based linear transformation of sb–t are near 0.2. In the case of tests with a strongly unbalanced answer key, the magnitude of pb can underestimate the tested ability range, to the point that it is effectively meaningless.

Table 4: Various indices for interpretation of discrimination index (power)

Discrimination power	**Interpretation**	**Author**
≥0.35	Excellent	Elfaki, Bahamdan et al.[41]
0.25–0.34	Good	
0.21–0.24	Acceptable	
≤ 0.20	Poor	
≥ 0.50	Very Good Item (Definitely Retain)	Obon and Rey[24]
0.40–0.49	Good Item (Very Usable)	
0.30–0.39	Fair Quality (Usable Item)	
0.20–0.29	Potentially Poor Item (Consider Revising)	

≤ 0.20	Potentially Very Poor (Possibly Revise Substantially or Discard)	
> 0.35	Excellent	Bhat and Prasad[32]
0.2–0.35	Good	
< 0.2	Poor	
>0.40	Very good	Sugianto[37]
0.30–0.39	Reasonably good possibly need to improvement	
0.20–0.29	Marginal item usually needing and being to improvement	
<0.19	Poor item rejected or improved by revision	
≥ 0.40	Very discriminating, very good item(Keep)	Aljehani, Pullishery et al.[42] and Sharma[8]
0.30–0.39	Discriminating item, good item (Keep)	
0.20–0.29	Moderately discriminating, fair item (Keep)	
< 0.20	Not discriminating item, marginal item (Revise/Discard)	
Negative	Worst/ defective item (Definitely Discard)	
> 0.30	Excellent discrimination	Ramzan, Imran et al.[43]
0.20–0.29	Good discrimination	
0–0.19	Poor discrimination	
00	Defective	
≥ 0.35	Excellent	Uddin et al.[35]
0.25–0.34	Good	
0.21–0.24	Acceptable	
< 0.20	Poor	

As a simple check of criterion-referenced equivalency, s²b coef is recommended. Departures from the C-R model can be detected by inspection of the variance p-d or pb-sq curves. A couple of retests may be needed to consider possibly poor items for additional adjustment or elimination. At this step, final item discrimination should be computed and compared with the construction specification. In general, all left screening tests (incompetency wise) may imply a need to review examinee population characteristics to maintain test validity. In particular, if this phenomenon recurred consistently across test administrations, it may indicate a need to redesign the assessment system (e.g., test type or version). A systematic left screening failure may be associated with test administration procreative workarounds such as, but not restricted to, item or answer key security breaches, rogue training, coaching or tutoring, cheating or fraud.

IV. DISTRACTOR ANALYSIS

A stem with or without a leading question and five or four choices are typical of Type A MCQs. One answer is crucial, while others are distracting.[8] Distractors should be convincing and misrepresent the core solution. The distractors should match the key answer in word usage, syntax, style, and length.(44) DE measures how bad responses distract students.[24] Functional distractor (FD) is a distraction chosen by 5% or more of the examinee.[8, 45] Less than 5% of examinees chose non-functional (NFD).(8) Other research found 1% of examinees were functional distractors.[46, 47] Items are typically categorized by NFDs.[21, 24, 38, 39]

Table 5: Classification of items according to non-functional distractors

Number of NFD	Percentage	Interpretation
3	0	Poor
2	33.3	Moderate
1	66.6	Good
0	100	Excellent

NFD makes the object easier to distinguish than FD distractors.[28, 39] A negative connection exists between non-functional distractors and dependability.(28) Two main causes of non-functional distractions exist. First, assess the item writer or composer's training and construction talents. The second issue is the content-distraction gap. NFDs can be reduced by writing and building training.(39) NFDs are also caused by low cognitive level, few distractor possibilities, and logic cues.[48] Students may recognize the distractor as the wrong one if they learn the knowledge. Since NFDs do not affect test measurement, they should be eliminated or replaced if they have no other source.[24] High-scoring examinees who frequently chose distractors over the key answer may have poor drafting, a misleading question, or double-keying.[24, 49] Three options are more practical than four, do not affect reliability, and do not drastically affect discrimination index.[21, 38, 39, 47] Equal distractors in all exam items are not psychometrically supported.[21, 50]. Content that generates realistic distractions should determine an item's alternatives.[45, 49] Reduced options/distractors can speed up test replies, increase content coverage, reduce composer workload, and improve criteria.[51]. Puthiaparampil et al. discovered less negative and positive associations between functional distractors and difficulty and discrimination indices.[46] A strong significant correlation existed between DIF and NFDs.[52] Many research reported no link between DE, difficulty index, and discrimination index.[19, 21, 49, 53] DE and other item analysis measures like Cronbach alpha were uncorrelated by LiconaChávez et al.[53] Other writers claim low DE lowers the difficulty index.[54]

V. ITEM ANALYSIS TECHNIQUES

The techniques utilized in item analysis are primarily classified into two approaches: classical test theory (CTT) and item response theory (IRT).[12] Classical test theory is the focus of this analysis. It is the prevailing model for test construction and has been widely applied to evaluate the quality of test items, specifically multiple-choice questions (MCQs). A CTT-based item analysis employs operational scores as input data for a pre-selected group of individuals. Each item and the test as a whole are evaluated according to a number of predetermined criteria and statistics. Descriptive statistics are calculated from raw operational scores. In evaluating the performance of items, raw scores are categorized according to a number of specified criteria. Each of the categories is assigned a predetermined value.

The situation in which data are categorized inherently leads to intervals that bear the assumptions of distinct attributes of individuals or test items. An essential characteristic of categorical nonparametric approaches is that it is impossible to statistically control the influence of certain score ranges in item analysis or to study the consequences of a differential impact due to data categorization in single studies. Tests with different structures are usually considered separately. Item response theory offers a different approach to data

analysis, in which the performance of both items and testees is modeled by specific random variable distributions, which are described by mathematical equations. A common assumption of all IRT models is uni-dimensionality. This means that a single underlying variable (trait, ability, or skill) can be used to explain the performance of each examinee on the set of items. The discrimination ability of item the i is defined as the item parameter ai in the one-parameter model.

1. Classical Test Theory

As one of the basic methods in item analysis, classical test theory (CTT) provides a set of concepts and methods that form the foundation for the evaluation of item or test reliability.[55] The reliability or precision of a test is an important issue that all test developers have to take into consideration in order to produce high-quality tests. The reliability coefficients derived from CTT are used to examine and verify whether the test is reliable. Since the reliability indexes yield evidence towards the reliability of the score, it is crucial to report them with scores from the test in order to interpret believes about the degree to which the test is a reasonable basis for making decisions. In general, a reliable test is one which will produce the same outcomes if it is administered a second time, provided the individual's performance has not changed, and that the conditions for taking the test remain the same.

The reliability of the test is not sufficient alone. The test should be able to distinguish between, in general terms, those who possess a robust knowledge of the subject and those who do not 1. This ability to discriminate is dependent upon the items in any given test. In terms of multiple-choice questions, those items whose performance is unaffected by knowledge of the subject or which are capable of inducing the same level of confidence between high- and low-achieving are deemed unproductive. Assessment of the reliability and discriminatory power of the test items can be achieved by several methods collectively known as item analysis. In addition to assessing reliability and discrimination, the item analysis also examines key measures such as the item difficulty level, discrimination index, and point biserial coefficient.

The item difficulty level is the most basic measure of item performance in assessment. Item difficulty is defined in terms of the proportion of correct responses: where "N" is the number of students who responded to the item, and "R" is the number of students who answered the item correctly. Items with extremely low or extremely high difficulty values (< 0.2 or >0.8) are candidates for revision because such difficulties imply that at least a large number of students guess the answer or that the item is too easy for the students, respectively. In either case, the test is unlikely to achieve greater discrimination if these types of items are retained. The item discrimination index is a measure of how well an item discriminates between high-achieving students and low-achieving students with respect to the underlying domain knowledge addressed in the question.

2. Item Response Theory

The application of item response theory (IRT) in item analysis is a systematic and iterative process, which can be broadly categorized into four stages: model selection and goodness of fit (GOF) evaluation, parameter estimation, and result evaluation, including item selection.[56] Within this framework, it is important to recognize the distinction among test analysis, which

refers to the overall, aggregate assessment of a test, such as reliability and validity; explanation modeling, which involves the analysis of a single test item; and simple IRT indicator calculation, which includes the computation of key indicator values relying solely on item characteristics.[57] Unlike the total-score analysis approach, which views a test as a single entity, IRT treats a test as a collection of individual responses and item characteristics. For the analysis of a psychological test containing and evaluating the quality of test items, IRT is the natural choice. This is because an IRT model establishes a mathematical relationship between latent traits possessed by respondents, items parameters associated with test items, and the probability of a particular response pattern observed. With this, both the characteristics of test items, as well as the abilities of respondents, can be inferred from aggregate test results relevant to individual responses.

Item response theory (IRT) modeling, including model selection and assessment of model fitting for a given test, is now mainstream in item analysis. IRT models have been widely applied not only in the area of educational testing and psychometrics, but also in other areas, such as survey response analysis, health data analysis, behavioral scoring analysis, and a variety of social science applications. Generalizing the logistic model through the introduction of one or more latent factors, differential approaches evaluated items according to the characteristics of an item pool (location and concentration of t, as well as the respective a or b parameters employed in the logistic models), with only the regression slopes being treated as parameters modeled on the item level.

VI. ITEM ANALYSIS IN PRACTICE

Within educational and psychological settings, item analysis is often a standard procedure. Multiple-choice tests are frequently analyzed to assess the functioning of test items. Such listing of statistics is then evaluated for individual items to determine if any are unreasonably "bad" and should be discarded or revised 1. Tests with bad items could, of course, never be reliable tests and are useless for predicting ability level in whatever is being measured. On the other hand, perfectly functioning test items do not insure reliable tests. A test with perfect items could be composed of only one item; ideally, a test should contain a great many items. But the greater the number of test items, all other things being equal, the more reliable the test would be. The purpose of this paper is to illustrate the use of item analysis on actual data and to discuss how the results are important in assessing and refining the selection of test items for such important objectives as is [(1)] determining how well test items are functioning; [(2)] eliminating bad items from tests in use; and [(3)] making items more effective by revising them.

1. Application of Item Analysis in Education Settings

Generally, an analysis of test items in education settings focuses on student choice patterns rather than on the questions of items. Indeed, the effects of item or question characteristics on examinee performance numbers are of interest and have attracted widespread attention from psychometricians.[11] The items must be evaluated carefully and continued post-hoc analyses performed to help understand student behavior. Numerous studies have documented the strong impact that test items can have on item analysis, such particular characteristics as difficulty, position, and hints.

Item analysis is a method of reviewing items on a test, both qualitatively and statistically, to ensure that each meets minimum quality-control criteria. The objective of qualitative and statistical review is to identify problematic items on the test. Items may be problematic due to being poorly written, unclear accompanying information, lack of a clear correct response, obvious distractors, or bias. Item analysis generally consists of item difficulty and item discrimination. Item difficulty is defined as the proportion of correct responses and is on a scale of 0.0 to 1.0. Items with extremely low or extremely high difficulty values do not discriminate between students. The first level of analysis reveals how difficult each item is, indicating which items need revision or which concepts need further discussion.

2. Item Analysis in Psychological Testing

Item Analysis means the evaluation of test items from the point of view of their effectiveness and appropriateness. The items can be in the format of an essay, short answer, true or false, or multiple-choice questions. Items of psychological tests are selected with great care but nevertheless maladjusted items creep in. The item analysis is resorted to in order to test the effectiveness of test items 1. In educational measurements, there are two main frames for studying a test, Classical Test Theory (CTT), and Item Response Theory (IRT). CTT relies on the positions and formulations of test items, not on the underlying psychological concept that measures the attribute concerned. IRT concentrate on the interaction between item and person performance on the item through examining the examinee's response or the probability of a correct response.

VII. ITEM ANALYSIS SOFTWARE'S

There are software tools available for conducting item analysis, including commercial products, licensed packages, free downloads, and free web-based tools. Of these tools, there are three that are recommended, each with a different focus. With respect to item analysis, some were primarily designed for conducting an item analysis. Examples include EXCEL, a spreadsheet, and IAP software, a free download designed specifically for item analysis. Other commercial test-item analysis programs are available for proprietary statistical package programs such as SPSS, SYSTAT, and SAS. Others are primarily designed for submitting a set of items for analysis by a central site. An example of this is the MC-QDA-easy, a commercial educational service that analyzes scores from constructed-response or qualitative test items in group administration. With respect to discipline, MC-MIA, MC-QDA, and MC-ANALYZE were designed specifically with science courses in mind for evaluating responses from multiple-choice test items constructed by the faculty. For producing classroom-ready tests derived from course question pools and randomly generating preparation exams, QuEX has been designed specifically for analytical questions and is available for a nominal fee.

There are numerous test-item analysis programs available commercially and from educational institutions (some free of charge) that can provide documentation and more complete descriptions of their statistical capabilities and/or approaches to item analysis. In selecting a program that works for a particular situation, the type of interface provided (spreadsheet versus programming) and whether or not it attempts to correct for guessing on the part of the tested should be considered. Each program also has its own strengths and weaknesses in terms of producing detailed data upon which decisions can be made and/or its ease of use. Questions can be easily drafted, submitted, and returned with tests scored automatically.

Computers can easily tabulate aggregate results for items without requiring staff time to do so. The bulk of the workload in evaluating a test to be analyzed can usually be taken over by program software.

1. Popular Software Tools

The movement towards objective, machine-scored tests for large classes has led to an explosive growth in the number of multiple-choice questions. Nevertheless, the techniques for writing, revising, and appraising such questions have lagged behind. This section summarizes commonly used and widely-available software tools for item analysis. While some of the methods and measures used in item analysis assessments need improvement, software tools that at least perform traditional items analysis are needed. There is also a need for tools that provide content validation of questions. Tests of students' understanding of key concepts, such as using data to distinguish between natural and human-made phenomena, mining for ore, or interpreting motion, are invaluable for evaluating the impact of instructional changes and for conducting research on student learning 1. There is also a parallel and equally growing need for technology-based inquiry tools to facilitate their implementation in the classroom. Item analysis of the questions making up these tests and instrumentational analysis of the tests themselves would provide a substantial contribution to test development. However, there are few widely available and easy-to-use computerized item analysis programs, and even fewer that perform full instrumentational analysis. Perhaps the ideal item analysis program could be created by assembling the best features of several currently available programs. Such a program would be a tremendous asset to educators and researchers involved in any kind of assessment.

2. Features and Functions

Most test item analysis software focuses on the "what" of analysis features and functions, rather than the "how"—that is, how to utilize the features to revise item test items. Depending on the experience and complexity of test items undertaken, different options and recommendations can be made for the user. The guide below describes a number of the most common test item software and examines their features. The goal is to provide insight into specific capabilities available, with the idea options increasing in sophistication from left to right. In addition, tools are available which can speed up and make easier the plight of analyzing, producing, and revising test items, especially with large item pools.

There remains an extensive opportunity to seek suitable software to improve the quality of test items within the education industry. Most assessment opportunities in further and higher education are within a multiple-choice question format. This includes the production of basic science, mathematics and statistics assessments, which are increasingly required to be computer-based. In-house test item analysis processing generally requires extensive training that slows the production and training of new test writers. Standard test item analysis software can be basic and simple, but still improve the quality of test items and is better than relying solely on paper test item generation and analysis, which is often difficult or impossible to replicate and is cumbersome to search query.

VIII. COMMON CHALLENGES IN ITEM ANALYSIS

The advent of computerized item analysis programs has enabled teachers to more easily conduct item analysis. Nevertheless, due to a number of reasons, there are still teachers who may not understand the concept of item analysis or how to conduct it. This is essentially a tragedy in the assessment process. Moreover, there are other teachers who know how to conduct it but encounter a number of problems in it. Again, this is regrettable because educating students without improving assessment measures used in the education process is not ensuring quality education. The goal of this paper, therefore, is to investigate common challenges in item analysis as these may lead to fear concerning its application among teachers and eventually prevent it from being used. Such fear, if exists, must be analyzed and remedied. One widespread difficulty teachers encounter in item analysis is with sample size. A test item should be administered to a minimum of 30 students to permit classical item analysis. However, a typical class in non-English speaking countries may contain only 15 students or even fewer in a given subject matter area. Administration of a test to a small class means multifold losses. Firstly, it means losing meaningful information for the designation of test reliability estimates. Such a loss should not be ignored especially if the class is expected to take similar tests in the future. Secondly, it means raising doubt as to the viability of standardized tests developed for use by larger populations if these tests do not yield desired test reliability estimates when administered to smaller ones. Therefore, many teachers operating in a small class setting may face a dilemma as to whether or not to use standardized tests. Another common challenge encountered in item analysis is the problem of item wording. When a certain number of students failed in an item, the educators' first reaction would be to scrutinize the item. Although item analysis software packages generally provide item statistics in terms of p-values and difficulty levels, one crucial piece of information cannot be obtained through statistical item analysis: i.e., whether the item is confusing. Item confusability due to ambiguous wording or interpretation of certain key phrases in an item stem can be a source of error in a student's response lethal to the item's quality for the student. Unfortunately, due to its qualitative nature, confusability is less tractable than other item properties. Therefore, consideration in wording items is crucial at the time of constructing an assessment instrument. Guessing behavior is also a common challenge, which is troublesome for educators designing multiple-choice questions to assess student understanding. The challenge can be approached from two angles. Firstly, what approaches can be used as antidotes? Educators can give more tests since the more items there are, the less chance a guess has of being correct. Confidence-weighted scoring may also reduce guessing by awarding full points for correct answers, zero points for no answer, and penalizing for incorrect answers. Secondly, what additional information would be helpful? Heightened guessing has effects not only on p-value but also on the item discrimination index. The change brought to both can supply useful information concerning the nature of the problem. The challenge of items experiencing limited discrimination is often encountered but seldom dealt with. If there are items that do not distinguish between students or even just older students or only recent ones, the item needs to be analyzed very carefully. It may be desirable to modify or delete the item and substitute it with something better, considering the context it was designed to test and the requisite knowledge concerning it of both types of students in question. The challenge of items experiencing bias is possible but usually not encountered unless the educator is dealing with culturally and linguistically diverse populations or trained students. Bias against students is manifested through the inclusion of items concerning the test philosophy of education or schooling that are pertinent only to the

socio-cultural background of some but not to others. Since such bias is less likely to exist for the locally produced tests, it is possible for many teachers to avoid it.

1. Sample Size Issues

In a nutshell, there are three reasons why it is important to consider sample size in item analysis studies. Most importantly, sufficient sample size is necessary to make certain that the conclusions drawn from results will be accurate. The larger the sample size, the more elaborate tests and models can be applied to the data without distorting results with respect to actual item functioning. Also, larger sample sizes can generally provide more certain estimates of the statistics that are calculated and lower standard errors 16. Finally, larger sample sizes also make it more likely that the data will exhibit the essential characteristics and dependencies of item scores being examined (at least within a level of sample size that is practical). This is pertinent as many of the techniques explored in the preceding chapters rely on input data meeting elementary assumptions about test items, stem answers, or latent traits.
Too large a sample size can provide results that may be statistically meaningful, but of questionable worth in practice such as computing item difficulty estimates to the fourth or fifth decimal. There are still sample size needs that are quite general in nature. These apply regardless of the details of the particular item analysis models or methods in use. In doing a specific item analysis study, it is critical to have a feel for the needs relative to these more general issues.

2. Ambiguous Item Wordings

A very common and easily dealt with difficulty arises from the ambiguous wording of items. Sometimes an item may be inadvertently or inadequately worded so that it can be interpreted in two or more different ways. Very often, the ambiguity lies in the choice of a pronoun which is used without referring to the noun on which it depends. Such ill-phrased items are the easiest to point out and would, therefore, be expected to be the first to be revised. The item seems reasonably unambiguous. In fact, ordinarily speaking, it is easy to think of someone considered to be cordial. In general, the item might seem poorly conceived for a test. Still, it shows how candidates might interpret an item quite differently.

3. Guessing Behavior

Test items are viewed as a discrete random variable based on the true-false model. While students cannot be expected to answer some of the test items correctly, they may respond to the items but be incorrect. Blind guessing attempts are followed by arbitrary guessing attempts. If the probabilities of blind guessing do not change from item to item, the meaning of construct differences across the items may have been lost due to irrelevant test conditions.[58] Guessing behavior, whether it be a blind guess or a partial knowledge guess, can adversely influence the result of an educational assessment (and any other tests relying on multiple-choice items) since it increases the chance of answering correctly (and incorrectly).

There are many approaches devised to address guessing behavior. In IRT (item response theory) modeling, guessing ability is either explicitly included in the model construction or neglected entirely. The second alternative treatment is commonly used within most of the traditional factor analysis approaches, which cluster the items into subsets. However, this

assumption about guessing indicates that items high in guessing may also be viewed as items that are poorly written. Nevertheless, it does not necessarily assist with the identification of the item clusters that reflect the dimensionality of the test. Guessing behavior can be treated in the model construction phase, the fitting phase, or the second phase.

4. Limited Item Discrimination

Limited item discrimination. A positive discrimination value indicates the item functioned in a desirable manner in relation to the test as a whole. When the item discrimination value is close to zero, or negative, the item is indicative of poor discrimination. In the case of zero or limited item discrimination, the item does nothing to increase the predictive value of a test. Negative item discrimination indicates that those who answered the test item correctly also tended to perform poorly on the total test score. Thus, such an item indexes performance in a manner inconsistent with the rest of the test items; a person who answers this item correctly is likely to be found in the lower scoring group relative to the test as a whole.

Analysis of item discrimination reveals whether concurrent tests, prescriptive tests, or formative pretests are working as poor items on a standardized test (zero discrimination) or in a manner opposite to that of the rest of the test (negative discrimination). It is therefore important to note the potential for zero-discriminating items to thwart the intended purpose of the test and/or its interpretation. Items with points-biserial correlations close to zero or negative are often regarded as suspect, and such items should be reviewed. Items are likely candidates for review if they demonstrate either a lack of a relationship with the total score (zero discrimination) or an inverse relationship with this total score (negative discrimination).

5. Item Bias

Item bias is defined as the lack of equivalence of a test instrument (test items) in the performance of test takers.(59) In simple terms, it may be said that item bias indicates that a test construct is equally valid for all test takers. Item bias comes into play when constructs are valid for one group of test takers and are not valid for another group as evidenced in different performance of the two groups on the same item even after controlling for ability (ANOVA). When an item discriminates unfairly against one test taker group or favorably to another test taker group according to their background, culture, dialect, race, sex, geographic location, and so forth, it has item bias. Bias in standardized tests due to the use of student background or demographic variables, like race or socioeconomic status, is not unusual.

Bad tests with test item bias could have a disastrous effect on important policy decisions. If the result of a test is used to make an important decision regarding a student such as admission to a competitive institution, then it is critically important that such a test be unbiased. Many psychometrician researchers have accepted the fact that biased tests produce misleading conclusions about the quality, productivity, achievement, or learning capabilities of educational systems, and hence remedial actions based on their conclusions would not solve the underlying problems.

IX. ETHICAL CONSIDERATIONS IN ITEM ANALYSIS

As the constructs of interest widen to encompass more content (e.g., large-scale high-stakes tests covering literacy, numeracy, science, and social studies) and more complex and higher-order skills (e.g., critical thinking), attention should now be focused on fairness and bias in test items as twice the issue of the constructs measured. These two inquiries—fairness and bias in test items and security and confidentiality—fall under the umbrella of ethical considerations. Ethics broadly encompasses moral principles governing the behavior of individuals or groups; in testing, it pertains to honesty and fairness in test development, administration, measurement, interpretation, and use. In test development, ethical standards primarily specify fairness, bias, security, and confidentiality, while fairness and bias take on wider and deeper meanings in test use and consequences.

Moving on to research conducted on the fairness and bias of large-scale high-stakes tests, large-scale assessments have sizable and beneficial consequences for individuals, educational institutions, teaching and learning, as well as the social system.[60] Some groups, such as girls or students from nonnative-speaking backgrounds, may be disadvantaged because of cultural, political, or economic factors affecting their equity in education and educational opportunities from the system. All individuals who take tests should receive equal measurement and treatment regardless of their gender, ethnicity, religion, socioeconomic status, or the schools they attend. The standard requirement for tests is that they should be unbiased measures of the constructs they intend to measure and fairly and equally treat all individuals taking the tests.

1. Fairness and Bias

Ensuring fairness and equity in the assessment of knowledge and capabilities is an ethical principle recognized as "social responsibility" throughout the test validation process.[61] This responsibility not only pertains to the conditions of test development, construction, and administration but extends to the types of inferences that can be drawn from test results. Based on the ethical principles of fairness and equity, the strengthening implications in the current validation models would further link to good practices concerning construct definition, selectivity bias at the test-taker end, false generalization, and statistical measures to mitigate test bias.

An item is said to be fair and unbiased for a group if the probability of a response to the item does not depend significantly on group membership when differences in the population parameter values are controlled. Further, a test is fair to a particular group if the test is free from bias for that group. The phenomenon of fairness is, broadly speaking, "the absence of bias" and the more specific phenomenon of test fairness generally hinges on the mathematical definitions of bias. Bias in the context of assessment practice can be conceptualized as systematic or consistent measurement error that differentially affects the test scores of a particular subgroup when compared with the scores of the test population in general.

X. PRACTICAL APPLICATIONS OF ITEM ANALYSIS IN MEDICAL EDUCATION

The scope of practice of item analysis also extends to wider applications in different settings, spanning a variety of domains. In medical education, it can be applied in different ways, providing useful practical applications in medical assessment settings.

Curricular Evaluation and Assessment: In such evaluations, items from a 5-year assessment period can be pooled. The overall assessment quality can be determined according to the general norms accepted in literature along with the departmental evaluation of curricular content. Significant differences can be detected in assessment quality and curricular content across individual courses. This contributes to the ongoing multiple-choice questions developing process. A thorough evaluation of item analysis in terms of item discrimination and item difficulty, distracters analysis, and the development of 'good items' can be used MCQs, which were developed using principles of content validation in basic and clinical subjects, were able to assess the higher-order cognitive skills of application and analysis or API/I level cases.

Assessment Development: In referring to sensitivity and effectiveness analysis, assessing basic science content knowledge, basic science individual MCQ validation can be done. This involves assessing item-wise parametric statistics in terms of item difficulty (p-value) and item discrimination. This can be done by item analysis where 'good items' can be identified and cola strengths can be examined. Such a practical approach can identify new test items of disparity in terms of technical quality.[36] In quantitative text analysis, basic science individual MCQ validation in terms of specificity analysis can be performed.

1. Curriculum Evaluation

Curriculum evaluation is another potential area for the application of item analysis. Various forces, both internal and external, operate in the medical education milieu where the curriculum constantly evolves. A particular new curriculum introduces innovations that, on the one hand, alter the educational environment. These innovations, inspiring curiosity and perhaps feelings of excitement, are often regarded as elusive. Formative and summative evaluations are conducted to ensure proper implementation, effectiveness, and efficiency of the new curriculum. It is widely accepted that evaluation is a necessary and integral component of curriculum planning and development. Without adequate appraisal of the curriculum and its continuously altering elements and facets, the potential efficacy of such changes cannot be realized.

Since such a new and innovative educational environment featuring change, alteration, and novelty furthers observation and questioning of the critically vital actor, the student, a natural and possible option for the implementors of this new curriculum is to investigate students' learning outcomes in this environment. Like prologues, epilogues, and sequels to a play, the students, as active participants in the education-attainment process, are constantly molded, reshaped, and inscribed or inscribed by the curriculum as they "sit" in it for years. Various instruments developed to probe students' learning outcomes are items test questions. Consequently, item analysis is applied in the pursuit of the evaluation of the new curriculum. Probing tests from different domains of knowledge are item analyzed in light of the new

curriculum, and a parallel analysis is conducted in light of the previous curriculum. The two student populations are equated at the beginning of the analysis by generating forms of knowledge to be investigated that were equally to be or have been acquired in either curriculum, thereby avoiding confounding differences in, for instance, students' school-preparatory academic abilities. In addition, novice and more expert faculties in the disciplines concerned conducted an active collaboration in the development of the item pool to be probed in the item analysis. Further, the curricular change investigated entailed a global and radical change from one educational milieu or system to another when the theoretical assumptions underlying both systems were antimonial. Various MICs were item analyzed, and the analysis explored changes in students' mastery thereof in line with the curricular change. As expected, students were found to have significantly lost mastery of items probing the MCs set out in the micro modular preclinical physiology curriculum and the corresponding MCs in the prior content syllabus preclinical physiology curriculum. Thus, item analysis serves to make visible whether learning outcomes explicable in light of a curricular innovation are found at all.

2. Assessment Development

Assessment development naturally progresses from general standard-setting processes to specific assessment development and analysis, including item analysis to assist in refinement of the assessment 3. Given the unique requirements of constructing assessments within the context of medical education, including the development and alignment of the test items to each objective and in consideration of the content maps, item analysis should be a continually evolving process. Nonetheless, performing some basic item analysis can help identify the most problematic items after each assessment and make refinements to them. This may include identifying items that were either too difficult or too easy to consistently improve those within the hardest range and flagging items that may not align with the intent of the objective. Additionally, patterns in item performance across different demographics can highlight faster/easier items and biases.[62] After these items have been flagged, there are a number of steps that could be taken to modify them to ensure better performance. For items with disproportionately low p-values, analyzing commonly recorded student responses can highlight alternative options, structural issues within the stem or design (e.g., double-barreled questions, unclear intent of the question), or if there is too broad of a scope within this item. On the contrary, if there are items with consistently high p-values, the item may be too easy and require modification or removal.

3. Improving Test Items for Revisions

Dedicated to improving test items through revisions, this section explores item analysis and how the insights gained from it can further be used in enhancing the quality and relevance of test items. On reviewing the medical assessments, many test items may have been perceived as poorly framed/worded, out of the current curriculum mapping, too difficult or too easy, which need to be revised carefully. Revision is a complex craft. The craft of reviewing and revising test items properly requires extensive expertise and should include a comprehensive analysis of test items, tests, and the purposes, needs, and perspectives of the different stakeholders involved. Key issues involved in an item analysis and revisions of medical assessments such as the pre-analysis test item examination, acceptability analysis, item scoring and discrimination, item high ease, out-of-scope items, item wording analysis,

review, and rewording approach, and acceptance of test items to be retained for consideration in the revised version of the test would all further be discussed. Such approaches taken would provide a clear item analysis and revisions of the medical assessments used for measuring knowledge regarding medical education needs, curriculum mapping, and the quality of the test items used.

With expertise in an item analysis, these could be furthered explored and an approach and guidelines could be undertaken for the revisions of test items. In general, the norms are composed of test items that are reasonably well written, clear, relevant, and that function adequately and equitably as perceived by the different subgroups of the population. Items shown to be invalid are usually not considered for negotiations where test development involves item adaptation.

4. Evaluating Test Quality

Considering the entire collection of test items, a broad view of overall quality is obtained through item analysis of the test itself. The overall quality of medical assessments is concerned with the test as an assessment instrument and with the test items that were, in fact, used. Most assessment strategies involve a series of tests, some used more than once, and the assessment implications of this design depend on the quality of the overall test. Paper format tests, especially those of multiple-choice items and other "selected response" questions, have the advantage of allowing some examination of the quality of the entire test itself in terms of the effectiveness of its items. Item analysis of the test itself provides four main sets of information. First, a frequency count of the number of items at given facility levels indicates which facility ranges the items in the test fall. This is often expressed in the form of a histogram that displays the facility levels of multiple-choice tests, which are generally desirable ranges. Secondly, a count of the correlation of scores to the item indicates whether the scoring formula of the item produces similar effects on the total scores to those of the other items in the test. Thirdly, the analysis of the distractors of multiple-choice items highlights their relative effectiveness. Fourthly, simple descriptive statistics such as the mean score and standard deviation are provided.

XI. FUTURE TRENDS IN ITEM ANALYSIS

This section summarizes thoughts about the future of item analysis, focusing predominately on trends that might be considered advances in technology, but also including innovation in approaches to item analysis. Hopefully, these thoughts will not just be speculative about the future, but will prompt others to consider what "could be" in their own work, the potential for advances that can be pursued, as well as consequences of the trends just described.

Technology is an ever-expanding source of practical and theoretical inquiry, development, and advancement in all fields, including the technology of testing. Computer-assisted technology has been developing and applied to testing for over a quarter century. Currently, computer technology is used in testing in four broad areas: 1) aides to test development, 2) test administration, 3) scoring tests, and 4) interpretation of test results. Scoring of tests and interpretation of test results are areas in which item analysis applications are widely used today either naively (for the easy use of statistics from computer programs) or poorly (manipulation of naively generated statistics to claim a test was properly used when it was

not according to test standards). There is little evidence representing current use or more novel applications of item analysis in the other two areas (aides to test development and test administration). However, computer-aided technology in testing will increase with the availability of cheap and portable personal computers, with powerful educational test development and item generation programs readily available. Computer technology has opened a whole new venue of thought, research, and experimentation about testing. Where computerized testing goes from here is very difficult to predict.

1. Advancements in Technology

Discussing the future, attention turns to possible advancements in technology that could affect item analysis. It's noted that if items were developed using newer types of technology, a new kind of analysis would need to be implemented. Substrates are the forms of technology used to structure the stimuli and responses. These technological advancements might include computerized images, audio and video, interactivity, new constructed response structures, or cold fusion. Cold fusion pertains to the items being presented by new formats such as paper gaming and internet gaming. The implications for the analysis process are discussed in terms of distinguishes. It is pointed out that the technology needs to be in place before items can be designed using such technology. Certainly, item design would be anticipated to follow item analysis, whereby items would be designed in the future in ways discussed.

Concerning the evaluation of test items, the currently used technology, tele-vision or computer technology, is seen as rudimentary by western style test item developers, with the hope that a future global market would be harmonious enough to share in the development of innovative designs and finer testing battery commodities. However, tele-vision technology would be too expensive, slow, and immobile as technology advances nations. As technology but not the design was difficult to envisage, it was easier to consider designs notwithstanding technology. In any case, the harmony needed does not have to be anywhere as grand as the industrial revolution, as items and designs already exist for use in expanding economies.

2. Innovative Approaches to Item Analysis

Innovative approaches to item analysis with a focus on novel suggestions and new ideas for methodologies and strategies that could be developed to enhance the accurate analysis of test items or new methods for evaluation. The emphasis is on potential novel approaches to the science of item analysis and methodologies are desired for both qualitative and quantitative innovative analysis of items. Suggestions could be based on practice, on research and development efforts with promising but untried ideas, or on theoretical suggestions or conjectures. Focusing on the next generation of item analysis has the potential for meaningful and far-reaching impact upon both K-12 and higher education testing and assessment practices. It is envisioned that novel suggestions for item analysis methodologies or overall approaches could radically change test design, test development, item scoring, and selection and monitoring test items in interesting and beneficial ways.

Innovative and novel item analysis suggestions could encompass range of diverse methodologies and approaches for the analysis and evaluation of item difficulty, discriminatory power, distractor effectiveness, and other indices utilizing qualitative

strategies such as fuzziness, data mining, branching algorithm, and latent trait modeling as well as quantitative strategies such as artificial intelligence, simulation, and choice models.

XII. CONCLUSION

The data from item analysis can either be used to just select a few items considered to be the best or can lead to a decision not to use a test for which a poor item analysis has been obtained. Item analysis can also help those who wish to go on constructing new tests by giving an indication of the type of items which work well. Those new to item analysis may find it useful to peruse the statistics first arrived at when this method of test evaluation was first used, and then choose the likely type of items now to be included. A test is a collection of test items or questions that the test takers answer to obtain a score which reflects their knowledge of the content domain measured by the test. An item or question on an educational test is a statement of a problem or concept that must be responded to. It is necessary that all items on a given test measure the same content and use the same standard of difficulty so that a valid interpretation of the test score is achieved.

The analysis of the items is aimed at determining how well a set of items functions both in terms of their psychometric properties and their content and educational relevancy. Item analysis is concerned with ensuring the relevance and effectiveness of test items. Content relevant 'wrong' responses as perceived by the test taker are deemed particularly important because they typically target specific misconceptions that point to where an item might be improved. On the positive side, an item with no content relevant 'wrong' responses might be too simply interpreted by test takers to discriminate between those who know and do not know the concept being measured. In some educational and psychological assessment systems, item analysis is used in a three-tiered approach. After a rigorous and comprehensive test is carefully administered by the examiners, the items are meticulously and systematically processed according to the three distinct tiers of analysis, each of which plays a pivotal and crucial role in the overall evaluation process. These tiers, meticulously designed and implemented, ensure that every item is thoroughly examined and evaluated, leaving no stone unturned in carefully scrutinizing the test performance. With their unique and distinct focuses, these tiers serve as the backbone of the entire analysis process, facilitating a comprehensive understanding of the test results and providing invaluable insights for further improvement and development.

REFERENCES

[1] Brown HD, Abeywickrama P. Language assessment: Principles and classroom practices: Pearson; 2019.
[2] Popham WJ. Test better, teach better: The instructional role of assessment: Ascd; 2003.
[3] Bachman LF, Palmer AS. Language testing in practice: Designing and developing useful language tests: Oxford University Press; 1996.
[4] Weir CJ. Language testing and validation. Palgrave Macmillan; 2005.
[5] Swanwick T, McKimm J. Assessment of leadership development in the medical undergraduate curriculum: a UK consensus statement. BMJ Leader. 2020:leader-2020-000229.
[6] Musial D. Foundations of meaningful educational assessment. (No Title). 2008.
[7] Brown JD, Hudson T. Criterion-referenced language testing: Cambridge University Press; 2002.
[8] Sharma LR. Analysis of difficulty index, discrimination index and distractor efficiency of multiple choice questions of speech sounds of English. International Research Journal of MMC (IRJMMC). 2021;2(1):15-28.
[9] Benson J. A Comparison of Three Types of Item Analysis in Test Development Using Classical and Latent Trait Methods. 1978.

[10] Kumar H, Rout S. Major tools and techniques in educational evaluation. Measurement and evaluation in education. 2016:256.
[11] Adegoke B. The role of item analysis in detecting and improving faulty physics objective test items. Journal of Education and practice. 2014;5(21):110-20.
[12] Ding L, Beichner R. Approaches to data analysis of multiple-choice questions. Physical Review Special Topics—Physics Education Research. 2009;5(2):020103.
[13] Odukoya JA. A critical assessment of the relationship between reliability and validity of some selected psychological tests. IFE PsychologIA: An International Journal. 1994;2(2):33-46.
[14] Singh T. Student assessment: Moving over to programmatic assessment. Medknow; 2016. p. 149-50.
[15] Salkind NJ. Encyclopedia of research design: sage; 2010.
[16] Green SB, Thompson MS. Structural equation modeling in clinical psychology research. Handbook of research methods in clinical psychology. 1989;138.
[17] Panayides P. Coefficient alpha: interpret with caution. Europe's Journal of Psychology. 2013;9(4).
[18] Tavakol M, Dennick R. Making sense of Cronbach's alpha. International journal of medical education. 2011;2:53.
[19] Mahjabeen W, Alam S, Hassan U, Zafar T, Butt R, Konain S, et al. Difficulty index, discrimination index and distractor efficiency in multiple choice questions. Annals of PIMS-Shaheed Zulfiqar Ali Bhutto Medical University. 2017;13(4):310-5.
[20] Rao C, Kishan Prasad H, Sajitha K, Permi H, Shetty J. Item analysis of multiple choice questions: Assessing an assessment tool in medical students. Int J Educ Psychol Res. 2016;2(4):201-4.
[21] Robinson JP, Shaver PR, Wrightsman LS. Measures of personality and social psychological attitudes: Measures of social psychological attitudes: Academic Press; 2013.
[22] Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychological assessment. 1994;6(4):284.
[23] Rezigalla AA. Item analysis: Concept and application. Medical education for the 21st century. 2022:1-16.
[24] Obon AM, Rey KAM, editors. Analysis of Multiple-Choice Questions (MCQs): Item and test statistics from the 2nd year nursing qualifying exam in a University in Cavite, Philippines. Abstract Proceedings International Scholars Conference; 2019.
[25] Hassan S, Hod R. Use of item analysis to improve the quality of single best answer multiple choice question in summative assessment of undergraduate medical students in Malaysia. Education in Medicine Journal. 2017;9(3).
[26] Cain MK, Zhang Z, Yuan K-H. Univariate and multivariate skewness and kurtosis for measuring nonnormality: Prevalence, influence and estimation. Behavior research methods. 2017;49:1716-35.
[27] Ali SH, Carr PA, Ruit KG. Validity and Reliability of Scores Obtained on Multiple-Choice Questions: Why Functioning Distractors Matter. Journal of the Scholarship of Teaching and Learning. 2016;16(1):1-14.
[28] Abdalla ME. What does item analysis tell us? Factors affecting the reliability of multiple choice questions (mcqs). Gezira Journal of Health Sciences. 2011;7(2):17-25.
[29] Vegada BN, Karelia BN, Pillai A. Reliability of four-response type multiple choice questions of pharmacology summative tests of II MBBS students. International Journal of Mathematics and Statistics Invention. 2014.
[30] Rosenberg SL. Multilevel validity: Assessing the validity of school-level inferences from student achievement test data: The University of North Carolina at Chapel Hill; 2009.
[31] Mehrens WA. 4. Assessing the Quality of Teacher Assessment Tests. 1990.
[32] Bhat SK, Prasad KH. Item analysis and optimizing multiple-choice questions for a viable question bank in ophthalmology: A cross-sectional study. Indian journal of ophthalmology. 2021;69(2):343-6.
[33] Chen J. Impact of instructional sensitivity on high-stakes achievement test items: A comparison of methods: University of Kansas; 2012.
[34] Cronje JH, Watson MB, Stroud L-A. Guidelines for the revision and use of revised psychological tests: A systematic review study. Europe's Journal of Psychology. 2022;18(3):293.
[35] Uddin I, Uddin I, Rehman IU, Siyar M, Mehboob U. Item analysis of multiple choice questions in pharmacology. Journal of Saidu Medical College, Swat. 2020;10(2).
[36] Kaur M, Singla S, Mahajan R. Item analysis of in use multiple choice questions in pharmacology. International Journal of Applied and Basic Medical Research. 2016;6(3):170-3.
[37] Sugianto A. Item analysis of English summative test: Efl teacher-made test. Indonesian EFL Research and Practices. 2020;1(1):35-54.

[38] Date AP, Borkar AS, Badwaik RT, Siddiqui RA, Shende TR, Dashputra AV. Item analysis as tool to validate multiple choice question bank in pharmacology. International Journal of Basic & Clinical Pharmacology. 2019;8(9):1999-2003.
[39] Kumar D, Jaipurkar R, Shekhar A, Sikri G, Srinivas V. Item analysis of multiple choice questions: A quality assurance test for an assessment tool. Medical Journal Armed Forces India. 2021;77:S85-S9.
[40] Earnest DS. Calculating item discrimination values using samples of examinee scores around real and anticipated cut scores: Effects on item discrimination, item selection, examination reliability, and classification decision consistency: The University of North Carolina at Chapel Hill; 2014.
[41] Elfaki OA, Bahamdan KA, Al-Humayed S. Evaluating the quality of multiple-choice questions used for final exams at the Department of Internal Medicine, College of Medicine, King Khalid University. Sudan Medical Monitor. 2015;10(4):123.
[42] Aljehani DK, Pullishery F, Osman OAE, Abuzenada BM. Relationship of text length of multiple-choice questions on item psychometric properties–A retrospective study. Saudi Journal for Health Sciences. 2020;9(2):84-7.
[43] Ramzan M, Imran SS, Bibi S, Khan KW, Maqsood I. Item analysis of multiple-choice questions at the department of community medicine, wah medical college, pakistan. Life and Science. 2020;1(2):4-.
[44] Considine J, Botti M, Thomas S. Design, format, validity and reliability of multiple choice questions for use in nursing research and education. Collegian. 2005;12(1):19-24.
[45] Tarrant M, Ware J, Mohammed AM. An assessment of functioning and non-functioning distractors in multiple-choice questions: a descriptive analysis. BMC medical education. 2009;9:1-8.
[46] Puthiaparampil T, Rahman M. How important is distractor efficiency for grading Best Answer Questions? BMC medical education. 2021;21:1-6.
[47] Gajjar S, Sharma R, Kumar P, Rana M. 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. 2014;39(1):17-20.
[48] Fozzard N, Pearson A, du Toit E, Naug H, Wen W, Peak IR. Analysis of MCQ and distractor use in a large first year Health Faculty Foundation Program: assessing the effects of changing from five to four options. BMC medical education. 2018;18:1-10.
[49] Sajjad M, Iltaf S, Khan RA. Nonfunctional distractor analysis: An indicator for quality of Multiple choice questions. Pakistan Journal of Medical Sciences. 2020;36(5):982.
[50] Haladyna TM, Downing SM, Rodriguez MC. A review of multiple-choice item-writing guidelines for classroom assessment. Applied measurement in education. 2002;15(3):309-33.
[51] Abdulghani HM, Ahmad F, Ponnamperuma GG, Khalil MS, Aldrees A. The relationship between non-functioning distractors and item difficulty of multiple choice questions: a descriptive analysis. Journal of Health Specialties. 2014;2(4):148.
[52] Alhummayani FM. Evaluation of the multiple-choice question item analysis of the sixth year undergraduate orthodontic tests at the faculty of dentistry, king abdulaziz university, saudi arabia. Egyptian Orthodontic Journal. 2020;57(June 2020):10-20.
[53] Velázquez-Liaño LR. Quality assessment of a multiple choice test through psychometric properties. 2020.
[54] Hassan S. Item analysis, reliability statistics and standard error of measurement to improve the quality and impact of multiple choice questions in undergraduate medical education in faculty of medicine at UNISZA. Malaysian Journal of Public Health Medicine. 2016:7-15.
[55] Thirakunkovit S. An evaluation of a post-entry test: An item analysis using Classical Test Theory (CTT): Purdue University; 2016.
[56] Nunes S, Oliveira T, Oliveira A, editors. Item response theory–A first approach. AIP Conference Proceedings; 2017: AIP Publishing.
[57] Chen Y, Li X, Liu J, Ying Z. Item Response Theory--A Statistical Framework for Educational and Psychological Measurement. arXiv preprint arXiv:210808604. 2021.
[58] Yeh C-C. The effect of guessing on assessing dimensionality in multiple-choice tests: A Monte Carlo study with application: University of Pittsburgh; 2007.
[59] Alordiah C, Agbajor H. Bias in test items and implication for national development. Journal of Education and Practice. 2014;5(9):10-3.
[60] Park YS, Yang EB. Three controversies over item disclosure in medical licensure examinations. Medical Education Online. 2015;20(1):28821.
[61] Banerjee HL. Test fairness in second language assessment. Studies in Applied Linguistics and TESOL. 2016;16(1).
[62] LaDuca A, Downing SM, Henzel TR. 5. Systematic Item Writing And Test Construction. 1995.