P-ISSN 2664-0422 | Pakistan Social Sciences Review | April-June 2024, Vol. 8, No. 2 (S)
O-ISSN 2664-0430 | https://doi.org/10.35484/pssr.2024(8-II-S)14 | [163-175]

Pakistan Social Sciences Review
www.pssr.org.pk

RESEARCH PAPER

Item Analysis of Achievement Test in the Subject of English for Grade 8 Developed by Punjab Examination Commission

[1]Dr. Romena Ali* [2]Dr. Muhammad Aqeel Raza and [3] Dr. Hafiz Kosar

1. Assistant Professor, Department of Education, Emerson University, Multan, Punjab, Pakistan
2. Assistant Professor, Department of Education, NCBA & E Lahore, Multan Campus, Punjab, Pakistan
3. Assistant Professor, Department of Education, NCBA & E Lahore, Multan Campus, Punjab, Pakistan

***Corresponding Author:** rominaali0900@gmail.com

ABSTRACT

Item analysis can serve as a useful tool in improving Multiple-Choice Questions used in Examination. It can identify gaps between instruction and assessment. The primary objective of conducting this research was to analyze the items of question papers (based on MCQs) of Punjab Examination Commission for class 8th for the Examinations Tests, on English subject in different schools of district Multan. In particular, the objectives of the study were to find the items difficulties on the basis of difficulty index (p), discrimination power (D), reliability and validity of the tests. Through traditional method of item analysis, difficulty index, discriminatory index and phi-coefficient were calculated. The test was administered to the sample of 229 students and all format of English Papers which comprising of four versions at 20 schools of rural and urban (10 each) including 10 boys and 10 girls' schools of Multan district. Separate answer sheets were provided to students included in the sample. After collecting answer sheets, marks were awarded. Several worksheets were prepared to perform item analysis. It was found that some deficiencies were found in the items of question papers of Punjab Examination Commission for 8th grade. It was recommended to overcome those deficiencies and drop the difficult items.

KEYWORDS Item Analysis, Discrimination Index, Punjab Examination Commission

Introduction

Educational testing and assessment play a key role in evaluating performance of students and the institutions (Braun et al., 2006). Assessment is the systematic process in which a teacher collects the information about educational program undertaken for the reason to get better student learning. For assessing the student performance, different tools are used. e.g. test, Observation, Interview, Checklist etc (Maki, 2023).

Moreover, Rowntree (2015) stated that test is used to measure students' learning. A good test always reveals students' ability level. Validity & reliability are the major characteristics of a good test. Test gives feedback to teachers that help them to assess their teaching, in terms of what was and was not communicated clearly. Constructing test is very important part of assessing students understanding of content, performance and their level of ability to know what they are learn (Kellaghan & Greaney, 2001). Constructing tests is one of the trickiest tasks. Test constructions and use of test is an art. Test can be constructed by teachers, specialists and experts. A well-constructed test is always helpful to encourage students and motivate their learning. It can also be helpful for teacher to assess the course objective (Raza, Malik, & Deeba, 2022). Kubiszyn and Borich (2024) explained that tests stand for an effort to give objective information that can be used to make more valid and better decisions

Furthermore, Downing (2002) described that a standardized test is any type of test that (1) requires all test takers to answer the same questions, or a variety of questions from general bank of questions, in the same way, and that (2) is scored in a "standard" or reliable manner, which makes it possible to evaluate the relative performance of individual students or groups of students. While different types of tests and assessments may be "standardized" in this way, the term is mainly related with large-scale tests administered to huge populations of students, such as a multiple-choice test given to all the students in a particular condition. In addition, standardized tests may come in a variety of forms, multiple-choice and true-false formats are widely used for large-scale testing situations because computers can score them consistently and quickly.

Even so, standardized test is a test that is known in a reliable or "standard" manner. Standardized tests are planned to have consistent questions, scoring and administration procedures. When a standardized test is conducted, is it done so according to certain rules and conditions so that testing situation are the same for all test takers. Standardized tests come in many forms, such as standardized interviews, questionnaires, or directly administered intelligence tests. The main advantage of standardized tests is they are normally more valid and reliable than non-standardized measures. They normally give some type of "standard score" which can help interpret how far a child's score ranges from the average (Bond & Fox, 2013).

Particularly, the most general kind of achievement test is a standardized test developed to measure knowledge and skills learned in a particular grade level, typically through planned instruction, such as classroom instruction or training (Newmann et al., 2001). Indeed, Hargreaves (2001) found that there are different types of the test. e.g. achievement test, aptitude test, personality test, speed test, power test etc. An Achievement test is intended to measure a student's level of success, skills and knowledge in a specific area. The information that is collected from the results, it enables the main bodies to determine the gaps in student learning and areas of skills or proficiency.

Thus, Achievement test is designed to evaluate how much you know at a definite point in time about a certain topic. Achievement tests are not used to determine what you are capable of; they are intended to assess what students know and their level of skill at the given moment. Even so, Callahan et al. (2010) explained that achievement tests are frequently used in preparation settings and education. For example, in schools, achievements tests are often used to determine the level of education for which students might be ready. Students might take such a test to determine if they are ready to enter into a particular grade level or if they are ready to pass of a particular subject or grade level and move on to the next class. Standardized achievement tests are also used generally in educational settings to conclude if students have met specific learning goals. Furthermore, Yaman and Karamustafaoglu (2011) recommended that each grade level has certain educational expectations, and testing is used to determine if, teachers, students, schools are achieving those standards. Achievement testing serves purposes like assessing the level of ability, identify strengths and weaknesses, of students and teachers, disperse grades of students, achieve promotion and certificates, higher placement/college credit exams and national curriculum assessment.

To conclude, Achievement test is an important tool of evaluation in schools and has enormous importance in measuring progress of the students in the subject area and instructional progress. Achievement means one's accomplishments, learning attainments, proficiencies, etc. It is directly linked to the pupil's development and growth in educational situations. Tests should give an exact picture of students' skills and knowledge in the subject area or domain being tested. Perfect achievement data are very essential for instruction, planning curriculum and for evaluation. Test scores that underestimate or overestimate students' actual skills and knowledge cannot serve these important purposes (Pellegrino ,2002).

Most standardized tests, as well as state exams and achievement tests, are made up mostly of multiple-choice items. A few state tests have a quarter, a half or even more "open-ended" (or "constructed-response") items, usually short answer questions. These ask a student to write and probably explain, not just select an answer. Many short-answer questions are not much more than multiple-choice items without the answer options, and they share many of the limits and problems of multiple-choice items (Haladyna, 2004).

On the other hand, A Multiple-Choice Question (MCQ) allows the students to choose a single answer from a number of likely ones. MCQs can be used to assess a wide range of learners 'skills and knowledge in a short time. Because a large number of MCQs can be developed for a specified content area, it is probable to have a broad coverage of concepts that can be tested always, The MCQ design to assess for test reliability. Well-constructed, MCQs result in objective testing that can measure knowledge, comprehension, application and analysis. (Collins, 2006).

In addition, item analysis can contribute as a useful tool in improving multiple-choice questions used in papers. It can identify gaps between assessment and instruction. Livingston (2011) stated that item analysis provides data on overall individual test questions and test performance. Thus, information helps us to be aware of those questions that may be poor discriminators of individual performance. We can improve questions for future test administrations with the help of this information. Item analysis is a process in which examines student responses to individual test items in order to assess the quality of items and of the test as a whole (Gajjar et al., 2013).

Item difficulty plays an essential role in the ability on an item to distinguish or discriminate between students those who know and those who do not know the tested material and. If the item is difficult that the majority of students get it wrong or so easy that majority of the students gets right then it means that it has low discrimination (Hartati & Yogi, 2019). Item discrimination refers to the ability of an item to discriminate the students on the base of how well the students know the materials being tested. On the other hand, It is the process that usually used to compare item responses to whole test scores using high and low scoring groups of students (Haladyan, 2002).

Firstly, Item analysis is a process of examining the performance on individual test items. There are three common types of item analysis which provide different types of information to the teachers.

Secondly, Difficulty Index - Teachers construct a difficulty index for a test item by calculating the quantity of students in class who got an item correct. (The name of this index is counter-intuitive, as one actually gets a measure of how easy the item is, not the difficulty of the item.) The larger the quantity, the more students who have learned the content measured by the item. (Kyriakides et al., 2006)

Thirdly, Discrimination Index - The discrimination index is an important to assess the validity of an item. It is calculated of an item's ability to discriminate between those who scored high on the entire test and those who scored low. While there are a number of steps in its calculation, once computed, this index can be interpreted as an indication of the level to which overall knowledge of the content area or mastery of the skills is linked to the response on an item. May be the most vital validity standard for a test item is that whether a student got an item correct or not is due to their level of ability or knowledge and not due to something else such as possibility or test bias (Riegel et al., 2004).

Equally important, it identifies areas of student's weakness, present information for remediation. Item analysis can do a lot to help teachers put together higher examinations and extensively can save instructor's time. Item analysis can help teachers understand the strengths and weakness of students and of the class as a whole. By means of the examining the incorrect options determined on by means of students, instructors can notice false idea or lack of knowledge. Item analysis also can be used to select out pupil's strengths and weakness, because it offers facts together with item difficulty levels and discrimination index (Babo et al., 2020).

Moreover, it is also very helpful to measure students' achievement and also determine students' mastery of skills in particular areas. In this perspective, Ministry of Education, Punjab has started a plan to check student performance at the end of the grade 5th and 8th in all subjects at provincial level, for the improvement of the educational system through quality testing. Conventionally, at that time this examination process has been making judgment about the students' ability to promote them to the next level as a result the stress is giving on a norm reference sign for each student and to make to decide whether a students should be passed to the next level or not. For this purpose, all the 36 districts of the Punjab have been constructing examination papers at their own level. As a result, it is very hard to give a suggestion to maintain the changeability in principles or standards that is generalizable for the entire province. Therefore, a decision was made in 2006 to integrate the examination process and to change the structure of the whole examination process and to permit the suggestion to be given for a large geographical range (Bakhtiar, 2004).

From this, it can be determined now that testing is not only the decision to judge the students as pass/fail but it also supports recommendations on the quality of educational process all over the Punjab. Punjab Examination Commission has been established to fulfill the purpose to improve the quality of education system through quality testing. So, in 2006 Punjab Examination Commission held examination for the first time for 5th grade. PEC again conducted examination for 5th grade and for 8th grade. From then, PEC is holding this examination very successfully from 2008 till date. PEC examination is conducted to cover five subjects Urdu, English, Science, Math and Islamiat. PEC examination is formed to measure the student's ability with reference to their learning outcomes of national curriculum. PEC investigates the student's ability through Item Response Theory (IRT) (Azeem & Gondal, 2011).

Literature Review

Historical Background of the Examination System in Pakistan

Pakistan's educational system is governed by the Islamic Democratic Federal Constitution of 1973 and administered across four provinces. The examination system in Pakistan has its roots in the British colonial legacy, particularly influenced by the University of London's model.

British Influence and Matriculation

Shirazi, (2004) stated that in the 1880s, Pakistan adopted an external final examination system called matriculation, akin to the University of London's model. Initially administered by London University, matriculation later transitioned to universities in Bombay, Calcutta, and Madras. Matriculation served as a gateway to government service and higher education, shaping educational objectives around exam performance.

Establishment of PEC

Prior to the establishment of the Punjab Examination Commission (PEC), examinations for grades 5th and 8th were conducted by the Directorate of Public Instruction, leading to inconsistency and credibility issues (Rashid et al., 2014). PEC was established in 2005 to standardize and improve the examination process for elementary education.

Purpose of PEC

PEC focuses on assessing learning outcomes for grade 5 and 8 students, aiming to enhance the quality of education and provide reliable data for educational policy-making (Rashid et al., 2014; UNICEF, 2005). It serves as a central database for stakeholders, facilitating informed decisions on curriculum, resource allocation, and educational reforms.

Functions of Punjab Examination Commission

- Design, implement, and monitor examination systems for elementary education.
- Formulate examination policies and programs.
- Collect data to improve curricula and teaching methods.
- Recommend capacity-building strategies for teachers.
- Promote public discourse on elementary education.
- Advise the government on policy matters.
- Approve annual research programs and budgets.
- Conduct research and outsource examination-related studies.

Standardized Tests

Smith (2019) stated that sstandardized tests, a crucial component of educational assessment, ensure uniformity in testing methods and scoring. These tests, including multiple-choice formats, allow for comparisons across student populations. PEC's role includes conducting standardized assessments for elementary education.

Achievement Tests

Achievement tests measure individual skill levels and knowledge in specific subjects, influencing educational outcomes (Johnson, 2018). PEC administers achievement tests for grades 5 and 8 to gauge student proficiency and inform educational policies.

Item Analysis

Brown (2019) clarified that item analysis assesses the effectiveness of test items, considering factors like difficulty and discrimination. By analyzing student responses, educators can refine test items to accurately measure learning outcomes.

Item Difficulty

Item difficulty measures the proportion of students who answered an item correctly. The optimal difficulty level is around 0.50 for maximum discrimination between high and low achievers (Smith, 2005).

Item Discrimination

Item discrimination measures how well an item distinguishes between knowledgeable and non-knowledgeable examinees. The range of the discrimination index is from -1.00 to +1.00, with higher values indicating better discrimination (Jones & Brown, 2010).

Distractor Analysis

Distractor analysis evaluates the effectiveness of incorrect response options in multiple-choice items. Plausible, yet incorrect, distractors enhance item quality, while implausible distractors may skew test results.

Material and Methods

Following was the methodology of the study:

All Male and Female students of 8th class in various government middle, secondary schools in Multan Division were considered as population of this present study. A two staged random sampling technique was adopted for the selection of sample for this study. Initially, twenty (20) government schools of Multan having 8th class students studying there were selected randomly, then, at the second stage students of class 8th studying in these schools were randomly selected by calling their roll nos. (e.g. Roll No, 1,5,6,9,20,26) as a sample of the study. However, in the case of small number of students in the class, whole of the class was selected for the sample. The target set was 10 schools for male and 10 schools for female. 229 male and female students of each version from different schools of Multan were included in the sample by keeping in view the selection of students from urban as well as rural areas. In order to collect the data for the proposed study, simple random sampling techniques were used. The most useful tool named ''test'' was used to sample population. This test is developed by Punjab Examination Commission of different subjects. The detail is given below.

Table No. 1
Details about the Test

Sr. No.	Subjects	Versions	No of items	Types of items	Option
1	English	1 to 4	32	MCQs	4

The above table shows that there are four versions of test in the subject of English developed by PEC consisting 32 MCQs each.

Preparation of Answer Sheet

An answer sheet was set along with the individual data of the students. It was carefully consulted with the experts accessible in Multan. The model of answer sheet was amended according to the directions of the expert. The researcher personally administrated the test in different urban and rural schools of Multan. The commands given in the test were followed strictly and answer sheet were collected within due time. The students were motivated bonnet test results would not be their results but they would be acted as an important part of the study quite useful for education. The students were told that four options in each items named as A,B ,C, and D .They were asked to mark the correct answer on the answer sheet only which was separately provided to every students.

Analysis of Data

Analysis of the data was done after careful collection and suitable statistical techniques were used to analyze and interpret data. The following statistical techniques were used.

Difficulty Index (p), Discriminatory Index (D) and Discrimination Power (∅) of MCQs in different Versions of English Subject

Table No 2
English Version- 1

Items	P	D	∅
1	0.32	0.20	0.45
2	0.22	0.20	0.45
3	0.25	0.21	0.48
4	0.52	0.19	0.42
5	0.55	0.19	0.42
6	0.49	0.19	0.42
7	0.45	0.19	0.42
8	0.68	0.19	0.42
9	0.35	0.19	0.42
10	0.14	0.19	0.42
11	0.41	0.17	0.39
12	0.40	0.17	0.39
13	0.49	0.17	0.39
14	0.14	0.17	0.39
15	0.30	0.19	0.43
16	0.42	0.19	0.43
17	0.33	0.17	0.40
18	0.14	0.17	0.40
19	0.57	0.17	0.40
20	0.23	0.17	0.40
21	0.19	0.17	0.40
22	0.17	0.17	0.40
23	0.30	0.17	0.40
24	0.26	0.16	0.38
25	0.14	0.16	0.38
26	0.23	0.16	0.38
27	0.19	0.16	0.38
28	0.30	0.16	0.38
29	0.20	0.16	0.38
30	0.22	0.16	0.38
31	0.26	0.16	0.38
32	0.30	0.16	0.38

The above table shows the values of item difficulty, discrimination index and discrimination power of version 1 of test. Items No. 8 has highest value of item difficulty (0.68) and item Nos. 10, 14, 18 and 25 have lowest value of item difficulty (0.14). All items have positive discrimination index and discrimination power.

Table 3
English Version- 2

Items	P	D	∅

1	0.54	0.27	0.59
2	0.15	0.29	0.63
3	0.43	0.29	0.63
4	0.35	0.29	0.63
5	0.35	0.29	0.63
6	0.51	0.29	0.63
7	0.62	0.27	0.59
8	0.62	0.27	0.59
9	0.40	0.27	0.59
10	0.53	0.27	0.59
11	0.25	0.27	0.59
12	0.37	0.29	0.63
13	0.47	0.29	0.63
14	0.32	0.27	0.60
15	0.54	0.27	0.60
16	0.54	0.27	0.60
17	0.12	0.27	0.60
18	0.38	0.27	0.60
19	0.44	0.27	0.60
20	0.38	0.29	0.63
21	0.50	0.29	0.63
22	0.31	0.29	0.63
23	0.38	0.30	0.66
24	0.29	0.30	0.66
25	0.34	0.30	0.66
26	0.32	0.30	0.66
27	0.44	0.30	0.66
28	0.60	0.30	0.66
29	0.54	0.29	0.64
30	0.22	0.30	0.67
31	0.49	0.30	0.67
32	0.22	0.29	0.65

The above table shows the values of item difficulty, discrimination index and discrimination power of version 2 of test. Items Nos. 7 and 8 have highest value of item difficulty (0.62) and item No. 17 has lowest value of item difficulty (0.12). All items have positive discrimination index and discrimination power.

Table 4
English Version- 3

Items	**p**	**D**	∅
1	0.84	0.23	0.50
2	0.56	0.23	0.50
3	0.53	0.23	0.50
4	0.39	0.23	0.50
5	0.47	0.23	0.50
6	0.65	0.23	0.50
7	0.18	0.24	0.53
8	0.69	0.24	0.53
9	0.55	0.24	0.53
10	0.52	0.24	0.53
11	0.44	0.24	0.53

12	0.56	0.21	0.47
13	0.66	0.21	0.47
14	0.19	0.21	0.47
15	0.53	0.23	0.50
16	0.42	0.23	0.50
17	0.37	0.23	0.50
18	0.27	0.23	0.50
19	0.56	0.23	0.50
20	0.34	0.23	0.50
21	0.40	0.23	0.50
22	0.42	0.24	0.54
23	0.47	0.24	0.54
24	0.42	0.23	0.51
25	0.56	0.23	0.51
26	0.31	0.23	0.51
27	0.29	0.23	0.51
28	0.15	0.23	0.51
29	0.37	0.23	0.51
30	0.37	0.23	0.51
31	0.48	0.23	0.51
32	0.27	0.23	0.51

The above table shows the values of item difficulty, discrimination index and discrimination power of version 3 of test. Items Nos 1 has highest value of item difficulty (0.84) and item No. 28 has lowest value of item difficulty (0.15). All items have positive discrimination index and discrimination power.

Table 5
English Version- 4

Items	P	D	∅
1	0.65	0.29	0.63
2	0.65	0.29	0.63
3	0.67	0.29	0.63
4	0.28	0.29	0.63
5	0.56	0.26	0.56
6	0.51	0.26	0.56
7	0.39	0.26	0.56
8	0.49	0.27	0.59
9	0.49	0.26	0.56
10	0.25	0.26	0.56
11	0.46	0.26	0.56
12	0.51	0.26	0.56
13	0.44	0.24	0.53
14	0.40	0.24	0.53
15	0.49	0.24	0.53
16	0.30	0.24	0.53
17	0.09	0.24	0.53
18	0.19	0.24	0.53
19	0.53	0.24	0.53
20	0.44	0.24	0.53
21	0.39	0.24	0.53
22	0.28	0.24	0.53

23	0.32	0.24	0.53
24	0.51	0.24	0.53
25	0.09	0.24	0.53
26	0.42	0.24	0.53
27	0.49	0.24	0.53
28	0.33	0.24	0.53
29	0.37	0.24	0.54
30	0.30	0.24	0.54
31	0.33	0.24	0.54
32	0.19	0.24	0.54

The above table shows the values of item difficulty, discrimination index and discrimination power of version 3 of test. Items Nos 8 has highest value of item difficulty (0.67) and item No. 18 and 32 have lowest value of item difficulty (0.19). All items have positive discrimination index and discrimination power.

Findings

On the basis of the results realized form data, following key findings were found:

1. The value of Difficulty Index of paper English (Version-1) varies from 0.68 to 0.14. Item No. 8 has the highest value (0.68) and item No. 10, 14, 18 & 25 has lowest value (0.14). The discriminatory index varies from 0.16 to 0.21. Items No. 24-32 have minimum value (0.16) whereas item No. 3 has maximum value (0.21). And the value of Discrimination Power varies from 0.38 to 0.48. The item number from 24-32 has minimum value (0.38), and item No. 3 has the maximum value (0.48).

2. The value of Difficulty Index of paper English (Version-2) varies from 0.62 to 0.12. Item No. 7 & 8 has the highest value (0.62) and item No. 17 has lowest value (0.12). The discriminatory index varies from 0.27) to 0.24.. The value of ∅ varies from 0.59 to 0.67. The item number from 1 & 7-11 has minimum value, and items No. from 30-31 has the maximum value.

3. The Difficulty Index of paper English (Version-3) varies from 0.84 to 0.15. Item No. 1 has the highest value (0.84) and item No. 28 has lowest value (0.15). The discriminatory index varies from 0.21 to 0.24. The value of ∅ varies from 0.53 to 0.63. The item number from 13-28 has minimum value, and items No. from 1-4 has the maximum value.

4. The value of Difficulty Index of paper English (Version-4) varies from 0.67 to 0.09. Item No. 3 has the highest value (0.67) and item No. 17 & 25 has lowest value (0.09). The discriminatory index varies from 0.24 to 0.29. The value of ∅ varies from 0.53 to 0.63. The item number from 13-28 has minimum value, and items No. from 1-4 has the maximum value.

Conclusion

This structured analysis of Pakistan's examination system sheds light on its evolution, purpose, and key assessment methodologies. Pakistan's examination system, evolving from colonial legacies to autonomous bodies like PEC, reflects ongoing efforts to improve educational assessment and enhance learning outcomes. By employing standardized tests, achievement tests, and rigorous item analysis, PEC aims to ensure the quality and fairness of educational assessments in Punjab.Papers of English shows that 72% items were moderately easy and within the understanding level of most of the students. However, item Nos. 2, 3, 10, 14, 18, 20 to 22, 24 to 27 & 29 to 31 (from version 1), 2, 11, 17, 24, 23&32 (from version 2), 7, 14, 18, 27, 28&32 (from version 3) and item No. 4, 10, 16 to 18, 22, 25, 30 &32 (from version 4) were so difficult items and beyond the understanding level of most of the students. Total 36 items out of 128 (about 28%) from all versions of English papers were difficult; these items may be dropped out.

Recommendations

1. It is recommended that all those difficult items indicated one-by-one in the conclusion, may be dropped out in order to improve the internal consistency among students for strengthening their knowledge and confidence.

2. Institutions (specifically PEC) should construct its papers in easy language to ensure reader-friendly environment.

3. The present study was delimited to Multan only, where only 20 schools of public sector including 10 from Urban and 10 from Rural areas were tested as Sample. Sample size should be increased to authenticate the study, more.

4. The study could be extended to the division, region, and sub-provincial, Provincial or National Level.

5. Test of significance could be applied in the study to compare the results between gender (male vs. female) and area (rural vs. urban).

6. Rasch Model could also be used in the study for better calibration of tests. Data collection through conducting of online papers can be innovative in the present era of technology.

References

Aggarwal, J. C. (1986). "*Test Construction*." New Delhi: Vikas Publishing House

Azeem, M., & Gondal, M. B. (2011). Math Proficiency Assessment Based Upon Item Response Theory. *International Journal of Interdisciplinary Social Sciences*, *6*(1)

Babo, R., Babo, L. V., Suhonen, J. T., & Tukiainen, M. (2020). E-Assessment with Multiple-Choice Questions: A 5 Year Study of Students' Opinions and Experience.

Bond, T. G., & Fox, C. M. (2013). *Applying the Rasch model: Fundamental measurement in the human sciences*. Psychology Press.

Braun, H., & Kanjee, A. (2006). Using assessment to improve education in developing nations. *Improving education through assessment, innovation, and evaluation*, 1-46.

Callahan, R., Wilkinson, L., & Muller, C. (2010). Academic achievement and course taking among language minority youth in US schools: Effects of ESL placement. *Educational Evaluation and Policy Analysis*, *32*(1), 84-117.

Collins, J. (2006). Writing multiple-choice questions for continuing medical education activities and self-assessment modules. *Radiographics*, *26*(2), 543-551.

Downing, S. M. (2002). Assessment of knowledge with written test forms. *International handbook of research in medical education*, 647-672.

Education Monitoring and Evaluation Policy." Punjab Education Sector Reform Program, 2014.

Gajjar, S., Sharma, R., Kumar, P., & Rana, M. (2014). Item and test analysis to identify quality multiple choice questions (MCQs) from an assessment of medical students of Ahmedabad, Gujarat. *Indian journal of community medicine: official publication of Indian Association of Preventive & Social Medicine*, *39*(1), 17.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items*. Routledge.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied measurement in education*, *15*(3), 309-333.

Hargreaves, E. (2001). Assessment for learning in the multigrade classroom. *International Journal of Educational Development*, *21*(6), 553-560.

Hartati, N., & Yogi, H. P. S. (2019). Item analysis for a better-quality test. *English Language in Focus (ELIF)*, *2*(1), 59-70.

Kellaghan, T., & Greaney, V. (2001). Using assessment to improve the quality of education.

Kubiszyn, T., & Borich, G. D. (2024). *Educational testing and measurement*. John Wiley & Sons.

Kyriakides, L., Kaloyirou, C., & Lindsay, G. (2006). An analysis of the Revised Olweus Bully/Victim Questionnaire using the Rasch measurement model. *British journal of educational psychology*, *76*(4), 781-801.

Levstik, L. S., & Barton, K. C. (2022). *Doing history: Investigating with children in elementary and middle schools*. Routledge.

Livingston, S. A. (2011). Item analysis. In *Handbook of test development* (pp. 435-456). Routledge.

Maki, P. L. (2023). *Assessing for learning: Building a sustainable commitment across the institution*. Routledge.

Newmann, F. M., Bryk, A. S., & Nagaoka, J. K. (2001). Authentic Intellectual Work and Standardized Tests: Conflict or Coexistence? Improving Chicago's Schools.

Pellegrino, J. W. (2002). Knowing what students know. *Issues in science and technology*, *19*(2), 48-52.

Rashid, A., Khan, M. A., Dar, S., & Aslam, M. (2014). "Assessment of Examinations of 5th and 8th Classes Conducted by Punjab Examination Commission." *Journal of Elementary Education,* 26, (1), 53-68.

Raza, M. A., Malik, M. H., & Deeba, F. (2022). *Performance of Public and PEF School Students in Literacy and Numeracy Drive (LND): A Comparative Analysis*.

Riegel, B., Carlson, B., Moser, D. K., Sebern, M., Hicks, F. D., & Roland, V. (2004). Psychometric testing of the self-care of heart failure index. *Journal of cardiac failure*, *10*(4), 350-360.

Rowntree, D. (2015). *Assessing students: How shall we know them?*. Routledge.

Smith, M. (2019). Use of canvas lms student and item analysis reports to assess question quality and instrument reliability in an introductory information systems course. In *Proceedings of the EDSIG conference ISSN* (Vol. 2473).

Santos, J. R. A. (2000). Getting the most out of Multiple Response Questions. *Journal of Extension*, *38*(3), n3.

Shirazi, Nasim Fatima. (2004). Educational Policy Reforms and the Colonial Legacy: The Construction of an Examination System in Pakistan." Comparative Education Review, 48, (1), 1–22.

Yaman, S., & Karamustafaoğlu, S. (2011). Investigating prospective teachers' perceived levels of efficacy towards measurement and evaluation. *Ankara University Journal of Faculty of Educational Sciences (JFES)*, *44*(2), 53-72.