MULTIPLE CHOICE ITEM WRITING AND ANALYSIS (A FRAMEWORK FOR ACTION)

Flexibility comes from having multiple choices; wisdom comes from having multiple perspectives.
Robert Dilts

Atilla ÖZDEMİR

1. Introduction

Human beings are acquainted with measurement and evaluation from the moment they fall into their mother's womb. In this period, the health status of the baby is measured and the results are evaluated with the available standard data. Thus, decisions are made regarding the development and health status of the baby. Although we do not realize it until our school years, measurement and evaluation processes that appear in every aspect of our lives become a natural part of our lives. In our daily life, we often make measurements and evaluations to make decisions. For example, in order to choose shoes, we evaluate the shoes that are suitable for our foot size, we need the width and height dimensions to build a football field, we use lights to provide traffic order, and in this process, we try to develop an efficient model by using time measures. We need measurement and evaluation processes in many different areas such as these. All these measurements made in different areas have their own measurement tools and units. These tools and units allow us to make our measurements directly, indirectly or in a derived way.

It can be said that the most important process in which we realize the effect of measurement and evaluation in our daily life is our school years. The most important reason for this is the use of the measurements obtained from the achievement tests applied to the students in the evaluation of course achievement. However, when considered in national standard tests that involve high-stakes tests (Kumandaş & Kutlu, 2010) used for transitioning between levels and continuing higher education, measurement and evaluation processes become an important agenda for all stakeholders in education. Since the purpose of all achievement tests is to measure psychological structures, their development has shown parallelism. In this sense, attempts to measure psychological constructs can be considered as approaching the subject from a much broader perspective.

While tests are generally used to get to know individuals (Cronbach, 1990), psychological tests are used to make standard measurements of an individual's intelligence, abilities, skills, attitudes, etc. (Anastasi, 1988). The use of psychological testing and testing programs is traced back to ancient China around 2200 BC (Cohen & Swerdlik, 2018; Janda, 1997; Popham, 1999). The applied tests included a number of difficult processes called imperial examinations, which were used for the selection of officers and civil servants. The similarities of those tests to modern tests in terms of application methods are striking. The simplest explanation for this situation is that the exam administration procedures and the psychometric properties developed by the Chinese were used as a basis for similar applications in countries such as France (1791), India (1833) and the USA (1883). In exam applications, processes similar to modern practices such as keeping the names of the candidates confidential, filling the papers by different coders so that the handwriting is unrecognizable, implementing exams in small rooms in special exam buildings under similar conditions for each candidate, and employing at least two independent evaluators when evaluating exam papers have been developed. (Bowman, 1989; Cohen & Swerdlik, 2018).

Tests that have been applied since ancient China can be considered in 3 different ways; the first is the purpose of the test, the second is the content of the test, and the third is how the test content is applied. The use of tests has become widespread in many areas, including clinical, counseling, geriatric, workplace and military purposes. In educational environments, it is still frequently used in the measurement of cognitive, affective and psychomotor areas such as intelligence, achievement, attitude, coordination skills. Moreover, it is seen that certain standards regarding test environments have been established from the past to the present and special exam buildings similar to today's buildings have been built (Bowman, 1989; Cohen & Swerdlik, 2018; Wainer, 1987).

It is seen that the test contents differed by shifting from verbal tests to written tests in the process. While the official exams were administered orally at the University of Bologna (1219), a similar practice was also used in the recruitment of students at the University of Oxford (1639) (Janda, 1997). In this sense, it can be stated that the establishment of examination systems in universities relatively fell behind. Horace Mann has made some criticisms about the limitations of oral exams (Madaus, 1988). In the mid-1800s, Horace Mann was the secretary of the Massachusetts State Board of Education and was aware of the increasing student enrollment due to immigration. He considered that it was important to establish a common public-school system so that the current system would allow incoming immigrants to be successful. For this, he argued that standardizing curriculum and instruction would reduce the difficulties faced by the growing student population (Smith, 2002). At Mann's direction, in 1845, the Boston School Committee used written essay exams instead of the oral exams to which students were accustomed (Rothman, 1995). Mann stated that the written exams should be applied, and thus, the subjective question selection problem that may arise in oral exams will be eliminated. Written exams had important advantages such as a standardized process for each student, being able to ask more questions, and reducing chance errors. Horace Mann believed that regular written examinations could be valuable tools for comparing the quality of teaching among schools (Caldwell & Courtis, 1925; Madaus & O'Dwyer, 1999). Thus, in the middle of the 19th century, the use of written exams for university entrance became widespread both in America and Europe.

In addition, towards the end of the same century, the fact that psychology became a separate discipline and studies of many scientists such as Darwin, Galton, Cattel, Herbart, Weber, Fechner, Wund, Locke in the fields of individual differences and psychophysics increased the importance of tests (Özgüven, 2007; Wainer et al, 2000). The influence of these studies, especially Galton's works, extends to the present day. Influenced by his cousin Charles Darwin's work 'On the Origin of Species', Galton wrote the book 'Hereditary Genius' and his ideas on the genetic transmission of intelligence have been influential in modern psychology, educational sciences and many other fields since then (Gillham, 2001). As Galton was never satisfied with the problem until he found methods that would express the problems he dealt with in numbers and analyze them statistically, his studies had an important place in the fields of measurement and statistics. Belgian statistician Adolph Quelet is the first person to use statistical method and the normal probability curve (bell curve) in social sciences (Schultz &

Schultz, 2011). In addition, with the work of Galton, the foundations of asking questions such as ‘Rating scales’ and ‘Questionnaire’, which are frequently used today, were laid. The widespread use of psychological tests increased gradually with the publication of the intelligence test developed by the French psychologists Alfred Binet and Dr. Th. Simon (1905) at the beginning of the 20th century (Kite, 1915). While all of the Binet intelligence tests and revisions were individual tests that could be applied to adults, after a while, a group of tests was prepared for school children’s norms by Pyle (1913) (Boake, 2002). In the World War I that took place in the same period, a commission was established by the American Psychological Association (APA) upon the request of the US army to divide its soldiers into various classes during the war. This commission asked Arthur Otis, who was also a student of Lewis Terman, who created the first group intelligence test, for help in dividing the soldiers into various classes. Arthur Otis prepared the ‘Army Alpha’ group tests for the literate and the ‘Army Beta’ group tests for the illiterate (Bell, 1921; DuBois, 1970; Gregory, 2015; Janda, 1997; Otis, 1920).

With these developments, from the beginning of the 20th century, formal education institutions at the level of primary education became widespread. As a result, emphasis has been placed on measuring individual achievements. As a consequence of the rapid rise of the American industry in this process, Frederick Winslow Taylor’s standardization-based structure, which prioritized the system in factories, was adopted by Edward Thorndike and formed the basis of school systems. On the basis of Thorndike’s structure, adopting the view that ‘quality is more important than equality’, he adapted Taylorism to the education system to distinguish superior students from those below average (Rose, 2016). Thorndike was also influenced by Francis Galton in his practices. Thus, Thorndike had a great influence on the development of achievement tests in the process, and in the ongoing process, testing studies were also handled by disciplines other than psychology. Especially in the field of statistics, along with Spearman’s studies in which he presented the basic principles of psychometry, many statistical algorithms have been put forward for the calculation and interpretation of the reliability coefficient (Aiken & Groth-Marnat, 2006).

In the early 1900s, it was seen that both statistical studies on test processes and different types of tests emerged in order to measure performances related to different skills (Burt, 1911; 1972; Gooddenough, 1926; Lowell, 1919; Porteus, 1915; Woodworth, 1910). With these developments, tests started to be applied collectively, not individually, and the use of multiple-choice tests became widespread with group tests. This was first implemented in the USA in 1901 in the university entrance exam, and in the ongoing process, a committee was established on this subject and an aptitude test called ‘Scholastic Aptitude Test’ (SAT) was developed and started to be used in 1926. The use of the test has become increasingly widespread and has begun to play an effective role not only in university entrance but also in granting scholarships (Wainer et al, 2015). Today, many universities in the United States use secondary school grades, scores from talent and achievement tests such as SAT (Scholastic Aptitude Test), GRE (Graduate Record Examination) and GMAT (Graduate Management Admission Test). While letters of recommendation are also considered in this process, some universities may also apply a separate selection exam in addition to these documents (Erdoğan, 2003).

A similar situation was experienced during the transition process to higher education in Turkey. As a result of the increase in the demand for higher education, the process started with a central examination in Ankara University in 1964 and the Interuniversity Selection and Placement Center (ISPC) was established in 1974, and the central examination system for student placement in all higher education institutions started. This center, which was established in 1982, took over the task of developing and implementing the central examination system under the name of Student Selection and Placement Centre (SSPC). The only thing that has not changed in the central exams since 1974 is that the exam questions are multiple-choice items. Although the use of multiple-choice items in tests provides great advantages, it is criticized for its negative effects on pre-university education (Eşme, 2014).

2. Historical Development Process of Multiple-Choice Tests

Although the development of psychological tests can be traced back to before Christ (Table 1), the use of the multiple-choice item format has managed to become one of the most valid and popular test formats for the evaluation of knowledge.

Table 1: Historical Development Process of Psychological Tests*

Year	Occurring Event
B.C. 2200	It is known that the proficiency test is applied in China. The civil servants of the empire were evaluated periodically.
B.C. 400	Plato advises people to work in jobs that are consistent with their skills and knowledge—an idea that will be echoed by psychologists, human resources professionals, and parents for ages.
1734	Christian von Wolf is the author of two books, Experimental Psychology (Psychologia Empirica, 1732) and Rational Psychology (Psychologia Rationalis, 1734), which envisioned psychology as a science. Von Wolf, a student of Gottfried Leibniz, put emphasis on the idea of perceptions below the level of awareness, which was Leibniz's thesis, thus predicting Freud's idea of the unconsciousness.
1829	English Philosopher James Mill, in his Analysis of the Phenomena of the Human Mind, suggested that intelligence consists of emotions and thoughts. Mill envisioned an approach called structuralism that aims to reveal the basic components that make up the mind in experimental psychology.
1845	Printed exams were used for the first time by the Boston School Committee under the guidance of educator Horace Mann.
1864	George Fisher, a British teacher, constructed a series of assessments of simple questions and answered the test questions as a guide for assessing students.
1869	Sir Francis Galton, cousin of Charles Darwin, published 'Hereditary Genius', which had an important place in (a) the claim that genius was hereditary, and (b) the use of the statistical technique, which Karl Pearson would later call correlation. Galton would later make many different contributions to measurement with his discoveries and innovations.
1890	American psychologist James McKeen Cattell coined the term mental testing in a publication. Cattell would go on to establish several publications, notably the Journal of Science and Psychology. In 1921 he formed the Psychology Association for the 'beneficial use of psychology'.
1892	Psychiatrist Emil Kraeplin, working with Wundt, planned a research in which he used a word association test. Also, in 1892, the American Psychological Association (APA) was formed with 31 members, mainly thanks to the efforts of its first president, G. Stanley Hall.
1904	Charles Spearman, a student of Wundt in Leipzig, began to lay the groundwork for the concept of test reliability. Spearman also began to establish the mathematical framework of factor analysis. E. L. Thorndike's 'An Introduction to the Theory of Mental and Social Measurements', the first essential test book on educational measures, was published.

1905	Alfred Binet and Theodore Simon published a 30-item ‘intelligence scale’, developed in Paris to help identify school-aged children with intellectual disabilities. The concept of measuring intelligence was accepted by readers around the world.
1908	A revision of the Binet-Simon intelligence test was published.
1910	In his article ‘Handwriting’, E. L. Thorndike developed the ‘Children's Handwriting Scale’, one of the first standardized tests to include arithmetic, handwriting, language, and spelling. This article contained 16 handwriting examples ranked by skill level.
1914	The World War I contributed greatly to test-enforcement studies, as thousands of soldiers had to be selected very quickly for mental function and emotional disposition. Alpha and Beta (first group intelligence tests), used as exams in the army, were structured and applied on newly recruited soldiers.
1916	After years of research, Lewis M. Terman at Stanford University published the Stanford Revision of the Binet-Simon Intelligence Scale. This American adaptation and revision of the test, first developed in France, would become commonly known as the Stanford-Binet.
1926	The Council of Higher Education supported the development of the Scholastic Aptitude Test (SAT) and applied the test for the first time.
1927	Carl Spearman published two-factor theory of intelligence in which he assumed the general intelligence factor (g) and specific components of general intelligence. Also, in 1927, German neurologist Kurt Goldstein began a study on brain-damaged soldiers in World War I. Many of these tests examined the ability to make inferences.
1931	L. L. Thurstone published Multi-Factor Analysis, a landmark study that had far more impact than statistical analysis; this had the effect of focusing research attention on cognitive abilities.
1939	While working at Bellevue Hospital in New York City, David Weschler introduced the Weschler-Bellevue Intelligence Test, which was developed to measure adult intelligence. This test would later be revised and become the Wechsler Adult Intelligence Test. Subsequently, additional Wechsler tests used for children and preschoolers would be developed and periodically revised.
1940	The World War II brought about the urgent need for a tool to be used to qualify for recruitment. In the same year, psychologist Starke R. Hathaway and psychiatrist/neurologist John Charnley McKinley published their first newspaper article on the test they were developing, now known as the Minnesota Multiphasic Personality Inventory.
1951	Lee Cronbach developed the alpha coefficient to measure the reliability of tests. Cronbach’s formula was a modification of the KR-20 (the twentieth formula of George Frederic Kuder and Marion Webster Richardson). Conceptually, Cronbach’s alpha calculated the mean of all possible split-half test correlations corrected by the Spearman-Brown formula.
1954	Swiss psychologist Jean Piaget published an original and influential study on cognitive development in children.
1970	The use of computers increased in the design, management, conclusion, analysis and evaluation of tests.
1971	It was decided to use tests in job applications (USA).
1980	Frederic M. Lord’s book, ‘Applications of Item Response Theory To Practical Testing Problems’ was published. This book became a pioneering work in the field, just like the earlier works in the field by American Psychometrist M. W. Richardson (1891-1965), Danish psychometrist Georg Rasch (1901-1980), and others.
1985	The Standards for Educational and Psychological Testing was published.
1998	An article by Anthony Greenwald et al. in the Journal of Personality and Social Psychology presented a methodology for measuring implicit cognition through the implicit-association test.

* Some selected events are given in chronological order. (Source: Aiken& Groth-Marnat, 2006; Cohen & Swerdlik, 2018)

Table 1 shows that the multiple-choice test format was first used during the World War 1 in the Army Alpha test, which was used by the US Army to classify 1.5 million soldiers for military purposes (Downing, 2006). Today, multiple-choice items are widely used in a variety of settings, including school tests, university exams, professional aptitude tests, and even TV quiz shows. Undoubtedly, the most important reason why multiple-choice tests have such a wide usage area is their objective evaluation (Baker, 2001).

From a historical perspective, the reason for the dependence on the use of traditional multiple-choice exams is clear. Professor Wood of Columbia University participated in a collaboration with IBM engineers in 1934 to develop a mechanical test scoring machine. The first model was developed by science teacher R. Johnson. “The developed machine adopts the logic of reliably reading the marking number of the graphite pencil, which conducts electricity at predetermined positions on a piece of paper, from an ammeter” (Kezer, 2013, p. 12). As a result of this, costs decreased due to the labor required in scoring other question types, at the same time, exam booklets were reused, and testing programs were developed with the ability to conduct exams for large groups at the same time. As a result of these developments, it has become inevitable to adhere to the multiple-choice question type in exams. For these reasons, multiple choice items are mostly preferred in high-risk exams/assessments (high-stakes tests/testing/assessment) (Kumandaş & Kutlu, 2010). High-risk exams are used for exams that have very important results for individuals, such as transitioning to a higher grade or higher education, and thus, causing anxiety (Başaran, 2005; Casbarro, 2004; Cizek, 2001; Orfield & Wald, 2000; Özer- Özkan & Turan, 2021; Resnick, 2004). In this sense, such exams as HSES/LGS (High School Entrance System), BPT/TYT (Basic Proficiency Test), FPT/AYT (Field Proficiency Test), PPSE/KPSS (Public Personnel Selection Exam), APPEEE/ALES (Academic Personnel and Postgraduate Education Entrance Exam), FLPT/YDS (Foreign Language Proficiency Test), MSE/TUS (Medical Specialization Exam) can be identified as high-risk exams. The most obvious common feature of these exams is that the exam questions in each of them consist of multiple-choice items. High-risk exams also significantly affect in-class assessments. In his study, Sınacı (2019) examined teacher-made exams held before and after TEOG (central joint exam for transition from primary to secondary education) and found that in some schools, teacher-made exams conducted before TEOG included questions and scoring that were not similar to TEOG’s, while teacher-made exams conducted after TEOG showed a higher correlation with TEOG. He stated that the placement scores obtained from the TEOG were not calculated fairly since the results of teacher-made exams were tried to be equated to TEOG results. In the study conducted by Özdemir et al. (2021), a total of 60 exam papers were examined in 4 different courses, including Turkish, Mathematics, Science and Social Studies courses. It was found that 833 (51%) of 1632 questions in total were prepared in multiple-choice format. As a result, it can be stated that teachers have to plan an exam-oriented course content and build their teaching on multiple-choice questions (Çetin & Ünsal, 2019).

Despite the widespread use of traditional multiple-choice items (MC), it has been observed that there are some limitations. Two of the most obvious ones are test wiseness and cheating. These are the limitations that negatively affect the psychometric properties of the test when there are measures other than the information that a test wants to measure. It is possible to present the advantages and disadvantages of traditional multiple-choice items as in Table 2.

Table 2: Advantages and Disadvantages of Traditional Multiple-Choice Items

Advantages	Disadvantages
High content validity.	Students can answer the questions correctly by chance.
Difficulty of the question can be modified by changing the options.	Since the answers are scored correctly or incorrectly, it is not possible to distinguish between students who have no knowledge about the subject and those who have partial knowledge.
Rating is objective. Thus, high reliability measurements can be made.	Since factors such as reading speed in questions with long question stems will be included in the measurement process, students with different reading speeds at the same knowledge level may get different scores.
With item analysis, detailed statistics on items and options can be obtained.	It is difficult to measure behaviors related to the higher levels of Bloom's taxonomy.
Can be applied to large groups as it can be read quickly with the help of machines.	Familiarity with the test and cheating are relatively easy compared to other types of tests.
Can be used at all educational levels.	Designing a good one is difficult and takes a long time.
Since it consists of many questions, it is a wide-ranging exam with the highest validity and reliability.	Adjusting the difficulty level of questions requires expertise.

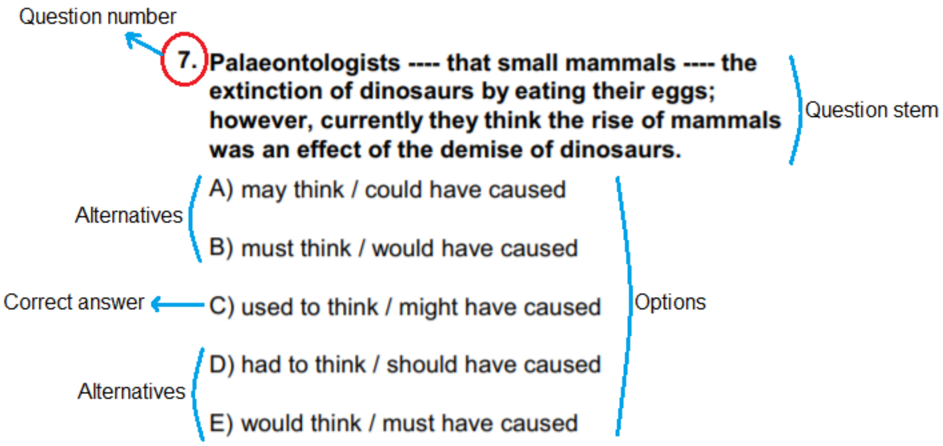
(Source: Atılgan vd., 2009; Baykul, 1999; Chatterji, 2003; Karataş et al., 2003; Kline; 2000; Kubiszyn & Borich, 1996; Miller et al., 2012; Özgüven, 2011; Murphy & Davidshofer, 2005; Turgut & Baykul, 2010)

Although there are various types of multiple-choice items, the most frequently used ones are traditional multiple-choice items. In addition, there are matched multiple choice items, alternative choices, best answer items, broad matched items, true-false items, complex multiple-choice items, multiple true-false items, content-dependent item sets, and discrete multiple-choice item types (Foster & Miller, 2009; Haladyna, 2004).

In the following section, the traditional multiple-choice item type, which is frequently used in large-scale and classroom measurements, is emphasized.

3. Writing Traditional Multiple-Choice Items

Traditional multiple-choice items basically consist of two parts: question stem and options. It is possible to detail these two parts as in Figure 1.



7. Palaeontologists ---- that small mammals ---- the extinction of dinosaurs by eating their eggs; however, currently they think the rise of mammals was an effect of the demise of dinosaurs.

A) may think / could have caused

B) must think / would have caused

C) used to think / might have caused

D) had to think / should have caused

E) would think / must have caused

(2021 YDT/ İNG)

Figure 1: Traditional Multiple-Choice Test Item Structure

Figure 1 demonstrates that the multiple-choice item consists of a question stem and options with one of them being the correct answer. In general, items are prepared with three (primary school), four (secondary school) and five (high school and above) options. The main purpose of this is to reduce chance factor, to distinguish students who do not have the expected information about the item or who do not know the correct answer.

Although the classical multiple-choice item seems like a simple structure, the usage areas of the tests in which these items are used are very wide. Such a situation reveals a number of basic features that should be considered for the preparation of multiple-choice tests. These features can be briefly listed as follows (Başol, 2013; Doğan, 2007; Güler, 2012; Haladyna, 1996; Haladyna et al., 2002; Otbiçer, 2004; Öncü, 2003; Özçelik, 1997; Turgut & Baykul, 2012):

1. The question stem should be prepared to measure a single outcome and should clearly state what the problem to be answered is.

2. Necessary information should be given only for those who have the measured gain in the question stem. Unnecessary, clue-like information and questions whose answers may vary from person to person should be avoided.

3. Each item should focus only on a single mental skill, rather than complex chains of behaviors and skills. Traps should be avoided.

4. The test items should be written in a clear language; the expression should be precise and should not allow for different interpretations. Words that will change the meaning of the sentence, such as 'often', 'sometimes', 'rarely', lead to different interpretations instead of giving certainty to the item. Provided that the item is kept clear, it should be tried to be expressed in a minimum of words.

5. The information given in the question stem must be scientifically correct and consistent, and must have a clear correct answer.

6. The difficulty level of the test items should be kept well below the average reading ability of the group that will answer the test.

7. Written sources such as test items and other written materials to be used in the tests, textbooks, etc. should not be used verbatim.

8. The test item should not ask for trivial details. Otherwise, perfect but equally invalid items can be created.

9. The items in the same test should not contain clues that will enable one other to be answered.

10. The options should be compatible with the question stem in terms of meaning and grammar. The lengths of the options should be close to each other.

11. Options should be similar in wording, length and scope. In terms of sentence structure, if the wrong options do not resemble the question stem or the correct option is more similar to the question stem, it will be easier for the respondents who do not have the expected information to select the correct one. The options should be independent of each other and one should not include the others.

12. In items where options can be put in order, their placement in the options must be in a certain order. For example, numeric options should be given in ascending or descending order.

13. The number of options should be appropriate for the level of the student to which the test is addressed. For example, three options in primary school; four options in secondary school; five options are recommended for high school and above. The number of options for all items in a test must be the same.

14. While writing the options, it is recommended to use capital letters so that they can be noticed more easily by the student. Also, the options should not be overlapping, and one option should not cover the other.

15. The options should not include ‘all’ and ‘none’; because these are attractive answers. Especially in questions asking ‘the most correct’ answer, ‘none’ should not be included in the options as this situation creates an expression disorder. In addition, in the questions with the ‘all’ option, the person who is sure that the two answers are correct will easily select the ‘all’ option.

16. Distractors should attract those who do not know and those who know wrong just in accordance with their purpose. Distractors that confuse the ones who know the correct answer should be avoided. Misconceptions should be considered in the writing of distractors.

17. The difficulty of the item increases as the degree of closeness of the distractors to the correct answer increases. Item difficulty is adjusted according to the purpose of the test. Therefore, the degree of closeness of the distractors to the correct answer should be adjusted according to the difficulty level of the test.

4. Calculation of Multiple-Choice Item Statistics

There are stages in the preparation of a test in which multiple choice items are used. Basically, the process of preparing a test consisting of multiple-choice items consists of 5 basic stages (Airasian, 1994; Anastasia, 1988; APA, 1999; Haladyna, 1997; 1999; Kline, 1986; Özçelik, 1997):

1. Determining the purpose of the test (identifying learning deficiencies, measuring the level of success, evaluating the curriculum etc.)

2. The scope of the test and the behaviors to be measured (searching for resources on the subject, seeking expert opinion, textbooks etc.)

3. Writing test items (There must be at least one item measuring each achievement in the final form, more than one item should be prepared in the trial form)

4. Item proofreading (whether the item is qualified to measure the behavior to be measured, whether there is a scientific mistake, whether it is understandable in terms of language, whether there are grammatical/spelling mistakes, whether the test and items are defective in terms of technical features, etc.)

5. Trial form (The materials to be used are selected)

6. Creating the form (Items are placed in the form according to the group level to be applied. At this stage, typesetting, font, etc. procedures are completed. Pilot test is conducted and the statistical analysis stage is started)

7. Analysis of the measurements of the test (The reliability of the scores obtained from the test is analyzed.)

8. Item analysis (Item discrimination, item difficulty and distractor efficiency.)

9. Creation of the final test (The final test form is prepared as a measurement tool based on the statistical results obtained from the pilot test.)

5. Psychometric Properties of Multiple-Choice Tests

In order for a test to be considered psychometrically adequate, it must be demonstrated by appropriate statistical methods that the test accurately measures the variable it aims to measure and that the results it gives are consistent (Cohen & Swerdlik, 2017). The psychometric values of the scores obtained from a multiple-choice test can be calculated by classical test theory (CTT) or item response theory (IRT). While calculations related to item response theory generally require large data (at least 500), classical test theory is preferred for small groups. In this study, calculations will be made with classical test theory.

The features that should be present in a test can be listed as reliability, validity and practicality.

Reliability: For a multiple-choice test to be reliable, its results must be consistently similar across different measures or on different raters. In addition, there should be consistency between the items of the test.

Internal Consistency: Internal consistency is an analysis that shows that test items consistently evaluate the same variable. In tests consisting of multiple-choice items and scored as 1-0 (true-false), one of the indicators of internal consistency is the statistically calculated KR-20-21 coefficients. The KR-20 coefficient and the Cronbach’s alpha value are equal. “The range of reliability measures are rated as follows: i) less than 0.50, the reliability is low, ii) between 0.50 and 0.80 the reliability is moderate and iii) greater than 0.80, the reliability is high” (Salvucci et al., 1997, p. 115). Another method that shows internal consistency is the correlation of the item score with the total score. Correlation of the items with the total score is expected to be 0.3 and above (Nunnally & Bernstein, 1994, p. 303). Similarly, the change in KR-20 value when the item is deleted is also important. If an item in a test is deleted and the internal consistency of the test increases significantly, then that item may be measuring a different variable than other items in the test (Field, 2003).

Factors to consider to enhance the reliability of a test:

1- As the number of questions increases, reliability increases.
2- Clearly understandable and answerable questions increase reliability.
3- Respondents should be encouraged to answer each question carefully and quickly.
4- The duration of the exam should be long enough for almost all students to answer all the questions.
5- Every exam should be scored in an objective way.
6- Difficulties should not be encountered during the implementation of the test.

Validity*:* Validity is a concept that describes how well a test measures its purpose. A valid test can accurately measure the target variable.

Content Validity: The concept of content validity is related to the inclusiveness of the tests regarding the features associated with the variable it aims to measure. The first step in content validity is to collect theoretical information about the subject while developing the test and to create a table of specifications accordingly.

Criterion Validity: Criterion validity is a concept of how useful the test result is. Criterion validity is very important as it is the practical demonstration of whether the test measures the variable it aims to measure.

Construct Validity: Construct validity is a concept related to whether the test items form a structure suitable for the theoretical knowledge on the subject. It is necessary for the test items to form sub-dimensions and a structure as a whole that is compatible with the relevant theory.

Face Validity: Face validity is related to the test taker’s ability to understand the characteristic that the test is intended to measure. Unlike other types of validity, it is determined by expert opinion rather than statistical calculations.

Factors to consider to enhance the validity of a test:

1- Each question should be prepared in such a way as to reveal and measure at least one of the behaviors we want to measure.
2- Each question should be prepared in a way that distinguishes those who have the target behavior we measure and those who do not.
3- After the question items are prepared, the opinions of the group teachers should be sought.
4- It is absolutely necessary to prepare the test reliably.

5- The test should be prepared in a way that it is both inclusive and representative of the curriculum.
6- The questions should be neither too hard nor too easy.
7- The same questions should not be used for successive years without changing them.
8- Any mistake that may occur in the scoring of the papers will also reduce the validity.
9- Cheating also reduces validity.

If the reliability of a test is low, its validity is also low. However, a test with high reliability may not have high validity, even may have low validity. Therefore, reliability is a prerequisite for validity, but it is not sufficient on its own.

Practicality*:* Practicality is the concept of how much benefit a test provides compared to its cost.

6. Item Analysis in Multiple-Choice Tests

In item analysis in multiple choice tests, three main topics will be focused on: item difficulty, item discrimination, and distractor efficiency.

Item Difficulty Index (p_j): The difficulty of a test item is the ratio of the number of correct answers to the number of all respondents (Özçelik, 1997, p. 123). Item difficulty index is the percentage of answering the question correctly (Tekin, 1984). Item difficulty index (pj) takes values between 0 (zero) and 1. The closer the item difficulty index is to 0, the more difficult the item (question) is, and the closer to 1, the easier the item (question). The item difficulty index of a medium difficulty item (question) is between 0.40 and 0.60, and nearly half of the respondents are expected to answer the question correctly.

n_j : the number of people who answered the item correctly,

n_s : total number of students,

p_j: The item difficulty value is calculated with the following formula:

$$p_j = \frac{n_j}{n_s}$$

Item Discrimination Index (r_{jx}): It is also called the item validity coefficient. It is the correlation of an item with the test. This statistic is calculated with one of the biserial or point-biserial correlation coefficients. The discrimination of a test item is its power to distinguish respondents who have probed behavior from those who do not (Özçelik, 1997 p. 123). In short, it is the power to distinguish between those who know the correct answer to the question and those who do not. A high item discrimination coefficient means that there is a high correlation between the item score and the test score, and that students who correctly answer that item get a high score in the entire test. If it is low, it means that students who score high on the whole test cannot answer that item correctly. Therefore, we obtain the knowledge of the degree of distinguishing between the student who knows and the student who does not know, with the power of item discrimination.

n_j: 27% upper group,

n_x:27% lower group,

n: The number of students who fall into the 27% lower or 27% upper,

r_{jx} :The discrimination index of the item is calculated with the following formula:

$$r_{jx} = \frac{n_{j-}\ n_x}{n}$$

Even though the item discrimination index is a correlation coefficient, we can analyze it together with the item difficulty and interpret it by using Table 3.

Table 3: Item Difficulty and Item Discrimination Index Ranges

Item Difficulty Index Value (p_j)	**Interpretation**	**Item Discrimination Index Value (r_{jx})**	**Interpretation**
0,00 – 0,19	Very Hard	-1,00 – 0,00	Item does not work
0,20 – 0,39	Hard	0,00 – 0,15	Very low discrimination
0,40 – 0,59	Medium Hard	0,16 – 0,29	Low level discrimination, should be corrected
0,60 – 0,79	Easy	0,30 – 0,49	Acceptable level of discrimination
0,80 – 1,00	Very Easy	0,50 – 1,00	Good discrimination

(Source: Kutlu & Vefikuluçay, 2003)

Distractor Efficiency: Analyzing the options is important to see which options work how well. For this, options are analyzed below based on real data. Analyzes were made using TAP (Test Analysis Program), which is a free program (Ayhan, 2010).

Video 1: Test Analysis Program (TAP)

7. Summary

This section focused on the basics of preparing a traditional multiple-choice test and analyzing test items. Additionally, information regarding how traditional multiple-choice items emerged in the historical process and their usage areas were provided. A sample video application about the reliability of the test results and the statistical analysis of the items were presented, and the data file was shared, allowing readers to use the application.

REFERENCES

Aiken, L, R. & Groth-Marnat, G. (2006). *Psychological testing and assessment (12th ed)*. Boston : Allyn and Bacon.

Airasian, P. W. (1994). *Classroom assessment (2nd ed.).* New York: McGraw-Hill.

American Psychological Association. (2000). The American Psychological Association Annual Report 1999. *American Psychologist, 55(8), 785–816.* *https://doi.org/10.1037/h0092931*

Anastasi, A. (1988). *Psychological testing* (6. edition). New York: Macmillan Publishing Company.

APA. (1999). *Standarts for educational and psychological testing*. The American

Atılgan, H., Kan, A. & Doğan, N. (2009). *Eğitimde ölçme ve değerlendirme [Measurement and evaluation in education]*. Anı Yayıncılık, Ankara.

Başaran, S. (2005). *Diğer ülkelerde lise bitirme sınavları ve Türk eğitim sistemi için lise bitirme sınavı önerisi [High school leaving exams in other countries and high school leaving exam recommendation for the Turkish education system]*. Millî Eğitim Bakanlığı Eğitimi Araştırma ve Geliştirme Dairesi Başkanlığı

Başol, G. (2013). *Eğitimde ölçme ve değerlendirme. (2. baskı) [Measurement and evaluation in education]*. Ankara: Pegem Akademi Yayıncılık.

Bell, J. C. (1921). Group tests of ıntelligence. an annotated list. *Journal of Educational Psychology, 12(2), 103–108.* *https://doi.org/10.1037/h0073682*

Boake, C. (2002). From the Binet–Simon to the Wechsler–Bellevue: tracing the history of ıntelligence testing. *Journal of Clinical and Experimental Neuropsychology, 24:3, 383-405, DOI: 10.1076/jcen.24.3.383.981*

Bowman, M. L. (1989). Testing individual differences in ancient China. *American Psychologist, 44(3), 576–578.* *https://doi.org/10.1037/0003-066X.44.3.576.b*

Burt, C. (1911). Experimental tests of higher mental processes and their relation to general intelligence. *Journal of Experimental Pedagogy, 1, 93–112.*

Burt, C. (1972). Inheritance of general intelligence. *American Psychologist, 27(3), 175–190.* *https://doi.org/10.1037/h0033789*.

Caldwell, O. & Courtis, S. (1925). *Then and now in education 1845-1923: A message of encouragement from the past to the present.* Yonkers-on-Hudson, NY: World Book.

Casbarro, J. (2004). Reducing anxiety in the era of high-stakes testing. *Principals, 83(5), 36-38.*

Cizek, G. J. (2001). Cheating to the test. *Education Matters Journal, 1(1), 40-47.*

Cohen, R. J., & Swerdlik, M. E. (2017). *Psychological testing and assessment (9th ed.).* McGraw-Hill Education.

Cohen, R. and Swerdlik, M., (2007). *Psychological testing and assessment: an introduction to tests and measurement (7th ed.)* McGraw-Hill.

Çetin, A., & Ünsal, S. (2019). Social, psychological effects of central examinations on teachers and their reflections on teachers' curriculum ımplementations. *Hacettepe University Journal of Education) 34(2): 304-323.* https://doi.org/10.16986/HUJE.2018040672

Doğan, N. (2007). *Çoktan seçmeli testler [Multiple choice tests]. (Ed. H. Atılgan) Eğitimde Ölçme ve Değerlendirme (5. Baskı) [Measurement and evaluation in education], (23-80).* Ankara: Anı Yayıncılık.

DuBois, P. H. (1970). Varieties of psychological test homogeneity. *American Psychologist, 25(6), 532–536.*

Erdoğan, İ. (2003). *Çağdaş eğitim sistemleri (5. baskı) [Contemporary education systems]*. İstanbul: Sistem Yayıncılık.

Eşme, İ. (2014). *Transition to higher education in Turkey*. Journal of Higher Education (Turkey). 4(3):148–157. doi:10.2399/yod.14.015

Field, A. (2013). *Discovering statistics using SPSS (4th. ed.)*. London: SAGE Publications Ltd.

Foster, D. F., & Miller, H. L., Jr. (2009). A new format for multiple-choice testing: Discrete option multiple- choice. Results from early studies. *Psychology Science Quarterly, 51, 355-369.*

Gillham, N. W. (2001). *A life of sir Francis Galton: from african exploration to the birth of eugenics*. England: Oxford University Press.

Goodenough, F.L. (1926). *Measurement of intelligence by drawings*. Yonkers-on-Hudson, N.Y., Chicago : World Book Company.

Gregory, R. J. (2015). *Psychological testing: History, principles, and applications (7th edition)*. Pearson Education Limited.

Güler, N. (2012). *Eğitimde ölçme ve değerlendirme [Measurement and evaluation in education]*. (4. baskı). Ankara: Pegem Akademi Yayıncılık.

Haladyna, T. M. (1996). *Developing and validating multiple-choice test items*. New Jersey: Lawrence erlbaum associates, publishers.

Haladyna, T. M. (1997). *Writing test items to evaluate higher order thinking*. USA: Allyn and Bacon.

Haladyna, T. M. (1999). *Developing and validating multiple-choice test items*. Mahwah, NJ: Lawrence
Erlbaum Associates, Inc

Haladyna, T.M. (2004). *Developing and validating multiple‐choice test items*. (Third Ed.), New Jersey: Lawrence Erlbaum Associates, Publishers.

Haladyna, T. M., Steven M. D. & Michael, C. R. (2002). "A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment." *Applied measurement in education, 15/3, ss.309-333.*

Haladyna, Thomas M.- Steven M. Downing- Michael C. Rodriguez (2002), "A review of multiple-choice

Janda, L. H. (1997). *Psychological testing: theory and applications (1st editio*n). Pearson.

Kezer, F. (2013). Comparison of the computerized adaptive testing strategies. (Unpublished dissertation). Ankara University. Ankara.

Kite, E. S. (1915). *The Binet-Simon measuring scale for ıntelligence: what it is; what it does; how it does it; with a brief biography of its authors, Alfred Binet and Dr. Thomas Simon*. Printed by the Committee on Provision for the Feeble-minded.

Kline, P. (1986). *A handbook of test construction*. New York: Methuen & Co. Ltd.

Kline, T. J. B. (2005). *Psychological testing: A practical approach to design and evaluation*. CA: Sage.

Kumandaş, H., & Kutlu, Ö. (2010). High stakes testing: does secondary education examination involve any risks? *Procedia- Social and Behavioral Sciences, 9, 758-764.*

Kutlu, Ö. ve Vefikuluçay, D. (2003). Development of reproductive knowledge test. *HIV/AIDS Dergisi, HATAM Yayınları, Cilt 5, Sayı 4.*

Lowell, F. (1919). A preliminary report of some group tests of general intelligence. *Journal of Educational Psychology, 10(7), 323–344.* https://doi.org/10.1037/h0071939

Madaus, G & O'Dwyer, L.(1999). A short history of performance assessment: Lessons learned. *Phi Delta Kappan, 80(9), 688-695.*

Madaus, G. (1988). *The influence of testing on the curriculum.* In L. N. Tanner (Ed.), Critical issues in curriculum: Eighty-seventh year-book of the National Society for the Study of Education (pp. 83-121). Chicago, IL: University of Chicago Press.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory (3rd Ed.).* New York: McGraw-Hill.

Orfield, G., & Wald, J. (2000). Testing, testing: The high-stakes testing mania hurts poor and minority students the most. The Nation, 270(22), 38-40.

Otbiçer, T. (2004). *Doğru sorular sorabiliyor muyuz? [Can we ask the right questions?]*. 13. Ulusal Eğitim Bilimleri Kurultayı, 256, 6-9 Temmuz, İnönü Üniv. Malatya.

Otis, A. S. (1920). The Selection of Mill Workers by Mental Tests. *Journal of Applied Psychology, 4(4), 339–341.* https://doi.org/10.1037/h0074893

Öncü, H. (2003). *Çoktan seçmeli testler [Multiple choice tests]*. TSA, 7(2), 87-103.

Özçelik, D. A. (1997). *Test hazırlama kılavuzu [Test preparation guide]*. Ankara: ÖSYM Yayınları.

Özçelik, D.A. (1992). *Eğitimde ölçme ve değerlendirme. (2. baskı) [Measurement and evaluation in education]*. Ankara: ÖSYM.

Özer-Özkan, Y. & Turan, S. (2021). More than meets the eye: the fact of the hıgh stake testıng. *Critical Reviews in Educational Sciences. Spring Cilt Volume 02 Sayı Issue 01 ss pp 59-63 ISSN - 2718-0808.* http://dx.doi.org/10.22596/cresjournal. 0201.59.63

Özgüven, İ. E. (2007). *Psikolojik testler (4. Baskı) [Psychological tests]*. Ankara: Nobel Yayın Dağıtım.

Popham, W.J. (1999). *Modern educational measurement: practical guidelines for educational leaders (3rd Edition).* Pearson.

Porteus, S.D. (1915). Mental tests for the feebleminded: A new series. *Journel of Psycho-Asthenics, 19, 200-213.*

Resnick, M. (2004). *The educated student: defining and advancing student achievement*. Alexandria VA: National School Boards Association.

Rose, T. (2016). *The end of average: how we succeed in a world that values sameness*. HarperOne Illustrated edition.

Rothman, R. (1995). *Measuring up: standards, assessment, and school reform*. San Francisco, CA: Jossey-Bass.

Salvucci, S., Walter, E., Conley, V., Fink, S., & Saba, M. (1997). *Measurement error studies at the National Center for Education Statistics (NCES).* Washington D. C.: U. S. Department of Education.

Schultz, D. P., & Schultz, S. E. (2011). *A history of modern psychology*. Cengage Learning.

Sınacı, B. (2019). *The comparison of scores in transition from basic education to secondary education (TEOG) and other scores*. [Unpublished master's thesis]. Hacettepe University. Ankara.

Smith, N. (2002). *American reading instruction (Special ed.).* Newark, DE: International Reading Association. (Original release date: 1934).

Tekin, H. (1984). *Eğitimde ölçme değerlendirme [Measurement and evaluation in education]*. Ankara:Has-soy Matbaacılık.

Turgut, M.F., ve Baykul, Y. (2012). *Ölçme ve değerlendirme [Measurement and evaluation]. (4. baskı).* Ankara: Pegem Akademi Yayıncılık.

Wainer, H. (1987). *The first four millennia of mental testing: from ancient China to the computer age*. Educational Testing Service, RR-87-34, Princeton. https://doi.org/10.1002/j.2330-8516.1987.tb00238.x

Wainer, H., Dorans, N. J., Eignor, D. R., Flaugher, R. L., Green, B. F., Mislevy, R. J., Steinberg, L., Thissen, D. (2015). *Computerized adaptive testing a primer (2nd edition).* Routledge. (Copyright Year 2000).

Woodworth, R.S. (1910). Race differences in mental traits. *Science, Feb 4;31(788):171-86. https://doi:10.1126/science.31.788.171*

To Cite this Chapter:

Özdemir, A. (2021). Multiple choice item writing and analysis (A framework for action). In Büyükkarcı, K. & Önal, A. (Eds.), *Essentials of applied linguistics and foreign language teaching: 21st century skills and classroom applications*, 190-208. ISRES Publishing.

ABOUT THE AUTHOR

Asst. Prof. Dr. Atilla ÖZDEMİR

ORCID ID: 0000-0003-4775-4435

atimaths06@gmail.com

Süleyman Demirel University

Dr. Atilla ÖZDEMİR graduated from Gazi University, Gazi Education Faculty, Primary Education Mathematics Teaching Program in 2005, Ankara, Turkey. In the same year, he started to work as a teacher at The Ministry of National Education. He completed his master's degree in Mathematics Education at Gazi University Primary Education Department in 2009, his master's degree in Educational Measurement and Evaluation in Hacettepe University Educational Sciences Department in 2014, and his doctorate in Gazi University Elementary Education Department Mathematics Education in 2016. He is an assistant professor of mathematics education at Suleyman Demirel University, Isparta, Turkey. He has participated in many national and international seminars, conferences and made oral presentations related to his field. In addition, he has written several book chapters and research articles. He also took part in national and international projects. He works on measurement and evaluation in education, technology integration in mathematics teaching, mathematical literacy in teaching and learning, and STEM education.