Testing & the Impact of Item Analysis in Improving Students' Performance in End-of-Year Final Exams

Aqeel Kadhom Hussein[1], Aqeel Mohsin Abbood Al-Hussein[2]

[1]The General Directorate of Education of Diwaniya, Iraq

[2]The General Directorate of Education of Babel, Iraq

Correspondence: Aqeel Kadhom Hussein, The General Directorate of Education of Diwaniya, Iraq.

Received: August 29, 2022 Accepted: October 15, 2022 Online Published: October 23, 2022

doi:10.5430/elr.v11n2p30 URL: https://doi.org/10.5430/elr.v11n2p30

Abstract

This research presented educational tests in a detailed way. Tests, their types, classifications and functionswere briefly discussed. Special focus was given to multiple-choice questions. The special design of this type of questions was explained and illustrated because it is the most important component of objective tests. The paper also presented a classification of tests that is based on their form and method of teaching. It concluded by stating four criteria of a good test. In addition, a recommendation to conduct further studies in this regard due to the huge importance of tests in general and their close connection to improving students' performance was also given. The paper provided suggestions for further studies in analyzing the quality of test items.

Keywords: testing, the impact, item analysis, improving, students' performance, final exams

1. Introduction

The purpose of this paper is to present testing as a very important component of the learning process. Also, it illustrates how the data obtained from testing in general can be used to improve and enhance both teaching and learning, in addition to improving the quality of test items. Gronlund and Linn (1990) state that tests' "data can be used to provide a basis for remedial work, to identify areas that need more extensive attention, and to suggest curricular revisions, or shifts in teaching emphasis". Woodford (1980) on the other hand, states that language tests were and still are used to assess "learners' knowledge of grammatical rules …" in addition to other purposes.

Many professionals in the field of education like Race et al.,(2005) believe that testing, or assessment plays an important role in deciding the effectiveness of the learning process through providing practical evidences. Test results are used to serve a number of purposes like: measuring the learning level, evaluating the method of teaching (pedagogy), deciding on what curricula should be implemented in a specific educational setting, in addition to comparing different educational systems (ibid).

1.1 Tests

In order to understand the nature of tests as a process, we need to consider the following definitions of tests and more specifically the nature of an educational test. M. Buchari states that a test is a "trial which is held to know some results from a certain subject which is taken from a student or a group of students (M. Buchari: 1987). W. James Popham (2020) in his article in ASCD educational magazine states that "an educational test is a formal attempt to determine a student's status with respect to specific variables, such as the student's knowledge, skills, and attitudes." There are different types of testing like: blood tests, geological tests and many other types of tests. However, our focus in this paper is not on those types of tests but on language testing and its different types in specific (Tim McNamara: 2000). The connection between teaching and testing is so obvious for both teachers and ordinary people alike. It is said that if "a teacher does a great instructional job, that teacher's students will usually perform better on tests" (W. James Popham: 2020). Also, the quality of teaching and the teaching methods should be reflected on the way "a teacher designs a test".

1.2 Why We Test

On one hand, tests in general are used to determine or achieve an accurate educational measurement that can lead to a good understanding of students' knowledge and learning abilities and levels. This understanding of the test philosophy

opens the horizon to a wider window to tests which is the types of decisions that can be affected by testing. Popham states that there are four different teaching decisions that are affected by tests and tests' quality, as follows:

- *Decisions about the nature and purpose of the curriculum*. Basically, every teacher tries to find good answers to the following questions:
    - "What am I really trying to teach?
    - What do my students need to know and be able to do?
    - How can I translate the big curricular goals set for my students into specific, teachable components?"

    Thus, tests can contribute to discovering the answers for all of these questions because they reflect the strong relationship between testing (as a process) and teaching.
- *Decisions about students' prior knowledge*. In order to figure out students' prior knowledge, teachers and educators need to seek answers to the following questions:
    - "What do my students already know about the topic I'm planning to teach?
    - Are there any gaps that I need to address before we can tackle this material?
    - Based on what my students know and can do, how can I tailor my instruction to provide the proper balance of remediation and challenge?"

    The answers to the above questions will determine what students already know about a specific topic, and this is going to facilitate teachers' mission to tailor some plans to avoid repeating already known pieces of information.
- *Decisions about how long to teach something*. The following set of questions shall be used to determine the duration of time required to teach each item:
    - "How long do I think it will take my students to master this content?
    - What kind of progress are they making?
    - Are we on the right track?
    - Should I continue teaching on my planned schedule, or are we ready to move on?"
- *Decisions about the effectiveness of instruction*. To achieve this purpose, teachers need to ask questions like:
    - "Did my students learn?
    - Was the instructional approach I took a good one?
    - What specific activities were the most advantageous?
    - Where do I need to make alterations?" (ibid)

Therefore, answers to the above questions help teachers find appropriate designs to their test items as well as these answers facilitate the mission for teachers to take the right decisions with confidence.

On the other hand, it is almost impossible to separate the function of test in education from the purpose of evaluation. Getting the verification data is considered as the purpose of evaluation. Besides that, it also can be used by all teachers and education supervisor to measure or asses the "effectiveness of teaching experience, learning activities and teaching methods" that are being used. (RiniYulia: 2006)

Anas Sudijono (year?) states that there are two functions for any test, as follows:

1. As a means of measuring the progress reached by a learner (or as known by an "educative participant") in a selected time.

2. As a means of measuring the learning program. Using a test that can measure "how far the instruction program has been achieved" is the best way to determine that. (Anas Sudijono: 2007).

Finally, Heaton believes that there are four different functions of any test, as follows:

> "1. To measure the students' performance in classroom.
>
> 2. To diagnose the students' weaknesses and difficulties.
>
> 3. To evaluate the effectiveness of a syllabus as well as the methods and materials used.

4. To provide the students with an opportunity to show their ability to recognize and produce correct form of language." (J.B. Heaton: 1979)

1.3 Types of Educational Tests

There are different major types of educational tests. They vary depending on their design or method and their purpose. According to their method there are two types: paper and pencil and performance tests. Whereas according to their purpose professionals in the field of educations believe that "the most familiar distinction in terms of test purpose is that between achievement and proficiency tests" (Tim McNamara: 2000). The most known forms of achievement tests n Iraq are end of year/course tests (summative tests), "observational procedures for recording progress on the basis of classroom work and participation" and monthly tests or formative tests (ibid).

Due to the strong connection between achievement tests and classroom instruction, they are considered apparently the most effective and widely used test. Therefore, more attention has been given to it in this paper. These tests "accumulate evidence during, or at the end of a course of study in order to see whether and where progress has been made in terms of the goals of learning… they support the teaching to which they relate" (ibid). Er.N.S (2012) believes that summative assessment in particular has a "substantial weight on students" that it is considered to be the main reason that forces them to learn. This type of testing can affect students and teachers at the same time as it "influence teachers' decision on what to teach in the first place" (ibid). regardless of what testing process is being followed, the validity – which is "the degree to which the inferences made on the basis of the assessment are meaningful, useful, and appropriate" (Wilson: 2007) of the testing tools must be verified to ensure the fact that "the inferred results are true indictors of students' knowledge and skills" (ibid).

However, some writers stand against using "multiple choice standardized tests" to perform achievements tests because they believe that "they have a negative effect on classrooms as teachers teach to the test" (ibid). This means that teachers prepare students to answer the test items and nothing more which is considered as a weakness to this type of tests as it is not directly linked to "language use in the world outside the classroom" (ibid).

There is another category for test types that mainly depends on the data obtained from pre and post-tests to determine whether or not the instruction along with intervention plans were useful of not as stated in Popham's article in the ASCD educational magazine. The first one "the pre-test" is mainly used to "to isolate the things your new students already know as well as the things you will need to teach them." (W. James Popham: 2020). Post-test data on the other hand can be used to make more defensible instructional decisions. (ibid)

Baily (1998) believes that tests can be classified according to their purposes into eight kinds of language assessments (tests): aptitude test, language dominance test, proficiency test, admission test, placement test, diagnostic test, progress test, and achievement test.

On the other hand, there is another classification of test types which is a little more detailed than the ones above. This classification of tests which can be used in many programs is built on what tests are based on meaning the purpose of tests; and classifies tests into five different types based on: organization, method of teaching, function, the form of the test, and scoring system (M. Buchari:1980).

1.3.1 Based on Organization

In terms of organization, tests are divided into standardized and teacher made ones. The researcher referred to the teacher made test as a theory.

a. Standardized Test

A standard test measures "the common objectives of a wide variety of schools have standard procedures of administration and scoring and provide norms for interpreting the scores." (E.N. Gronlund: 1981).

b. Teacher Made Test

This term refers to tests made by the teacher to achieve one specific purpose which is assessing the student learning. Due to that fact, it is valid in one place and is suitable for one setting. For example, it is considered valid in school A but not valid for school B. sometimes it is designed to assess students in one specific class, so it is only valid for this class. (Slameto E. Pendidikan: 2001)

1.3.2 Based on Method of Teaching

Based on method of teaching, tests are divided into two main categories: norm-referenced and criterion refence tests. Both types of tests will be briefly referred to in the following lines.

a. Norm-Referenced Test

E.N Gronlund states that norm-referenced tests (NRTs) are designed to rank or put pupils/students according to their level of achievement, arranging them from highest in performance to the lowest. This makes any achievement related decision like: selection, grouping or grading more confident which means that NRTs are "made to compare test takers to each other". (E.N Gronlund:1981).

b. Criterion Reference Test

A criterion-referenced test refers to a group of items that measure behaviors that are expressed in a set in a direct way (M. Ngalim Purwanto :2000)

1.3.3 Based on Its Function

Gronlund suggests four different categories for tests based on their function. The four types are: "placement, diagnostic, formative and summative tests." (E.N Gronlund:1981). In the lines below, each one of them will be briefly highlighted.

a) A placement test refers to a test that is designed to measure students' basic abilities in order to place them in the right or suitable level according to their ability.

b) A diagnostic Test is a test that is used during the teaching learning process to identify students' difficulties and weaknesses.

c) A formative test controls and determines the progress of learning. It also provides feedback for reinforcing the learning process, correcting learning errors and suggesting suitable intervention plans.

d) A summative test refers to the test given at the end of a course (in our case it is sometimes called end of year test). The results are primarily used for some purposes like assigning grades, determining students' mastery of the "instructional objectives".(ibid) Tinanunan in his book *Evaluation of Students Achievement* suggests another definition for summative tests based on the intention to show the standard reached by students now in comparison to other students at the same stage level. Hence, summative tests typically come at the end of a course or a unit of instruction. (Wilmar Tinambunan: 1998)

1.3.4 Based on the Form of the Test

Based on the form, there are two main kinds written tests and oral ones.

a. Written Tests

A written test refers to a test in which the items and the answers given to students are all written. In other words, testers have to answer all questions in the test by writing the answers on a paper known as the answer sheet (C. Thoha: 2003).

b. Oral Test

An oral test refers to the process of asking oral questions by teachers and students are supposed to answer these questions orally at the same time. J.B Heaton defines it as "testing the ability to speak." (J.B. Heaton: 1974)

1.3.5 Based on the Scoring System

Tests are divided into subjective test and objective tests based on the way they are scored.

a. Subjective Test

A subjective test is a test that depends on testers' personal evaluation of a particular test. It includes questions like essay writing, explaining sentences, interpreting, comparing through finding similarities and differences, clarifying and giving reasons or justifications for something. Subjective tests normally start with words like:"Explain…", "Clarify…", "Mention…", "Why…", "How…" (Anas Suduijono: 2005)

b. Objective Test

An objective test is a highly organized one which requires the pupils/students to fill in blanks, through providing a word or two, or to select the correct answer from a provided set of choices (Wilmar Tinambunan: 1988). It can be concluded that objective tests leave no space for testers to interfere with the test results, and that it is all the responsibility of test takers to decide on how well to do a test because most of objective test questions are in the form of multiple – choice ones (ibid). In the lines below, some examples of test items of objective tests will be highlighted.

1. Transformation

Ali is a very good writer. He writes ….. (a. goodly b. well)

2. Completion

It is half …… nine. (past / to)

3. Combination

I was running. The bell rang. (Join: use "when")

4. Addition (grammar)

Have you ……. seen Tom Cruise latest movie? (yet / ever)

5. Rearrangement

her – lost – he – market – the – in -, (Reorder)

6. Correct/incorrect (true/false)

Put a tick if the statement is correct and a cross if it is incorrect. (J.B Heaton: 1987)

Many of the above test items can be used to test vocabulary, reading comprehension, writing, listening, grammar and speaking skills (ibid).

7. Multiple-Choice Test

This particular test item is considered as the most important component of objective tests. Its techniques or principals are so important and must be taken into consideration when writing a multiple-choice test. Also, this test item is considered as one of the most useful of all objective test item types because it is easy to construct and simple to score and administer (Rahmadi Nirwanto: 2012). Rahmadi Nirwanto lists six principals that must be considered when constructing a multiple-choice question, as follows:

> "1. Each multiple-choice should have only one answer.
>
> 2. Only one feature at a time should be tested.
>
> 3. Each option should be grammatically correct when placed in the stem, except of source in the case of specific grammar test items.
>
> 4. All multiple-choice items should be at a level appropriate to the linguistic ability of test takers.
>
> 5. Multiple-choice items should be as brief and as clear as possible (though it is often desirable to provide short contexts for grammar item).
>
> 6. In many tests, items are generally arranged in rough order of increasing difficulty." (ibid)

2. The Design of a Multiple-Choice Test

Generally, a multiple-choice test has four options: (a), (b), (c), and (d). Only one of them represents the correct answer; this is known as the "key answer". The remaining three options are incorrect ones. These are known as "distractors".

These distractors should be plausible and should appear right to test takers who are unsure of the correct answer.

Heaton states that multiple-choice distracters should not be too complicated or requires high proficiency in language because too difficult options result in distracting good students more than students with low proficiency levels. (JB. Heaton: 1987)

Inspecting the way each multiple-choice item function is another important step in the analysis of multiple-choice items because it is important in criterion referenced test. (David P. Harris: 1969)

In order to determine the distracters effectiveness, we need to compare:

> "students' numbers on upper and lower group who choose the wrong alternative distracters. If an item contains distracters which attracted no one, not even the poorest examines, it is a nonfunctioning choice which will increase the chances that some examines will get the item right by guessing between or among the remaining two or three possibilities. The effectiveness of distracter defines by checking the frequency with which each distracter is selected by those failing an item." (ibid)

Thus, a distracter effectiveness is determined by how many times it is being selected by students.

Criteria of Good Test

In this part of the research, four criteria that determine whether a test is good or not will be briefly discussed. The four criteria are: validity, reliability, index difficulty and distracters.

2.1 Validity

Validity refers to "what characteristic the test measures and how well the test measures that characteristic" (HR-Guide: 2018). Also, test validity refers to the extent to which a "test measures what it is supposed to measure and nothing else" (J.B. Heaton: 1998). Heaton states that there are four types of test validity:

a. Face validity

The first of the four types is face validity which refers to the appearance of a test; in other words to the way it looks to test takers, testers, moderators and administrators. Based on this, it is useful to collaborate with other colleagues in writing a test; or at least showing it to them to figure out any potential absurdities or ambiguities of a test.

b. Content validity

The second type is known as content validity which mainly related to what kind of materials that students have learned. In this case, a good test should cover samples of the materials taught or to the teaching syllabus. Also, content validity depends on a careful analysis of the objectives of a course; meaning that it should be well constructed to be inclusive of the syllabus of a course. ((ibid))

c. Construct validity

A construct validity, which is the third type of validity, deals with test structure in accordance with a theory of language, behavior and learning. This makes a test capable of measuring specific learning characteristics. (ibid)

d. Empirical validity

The last type of validity is the empirical one. This type has two kinds: Concurrent and predictive validity. This classification is based on correlating test scores with subsequent or concurrent "criterion measures". For example, predictive validity of a test is determined when the scores of an English test as a foreign language are used to screen the level of university applicants and then these scores are correlated with students' grades in the first semester. Whereas the concurrent validity is achieved when a test is immediately followed up by rating each student's English proficiency depending on his/her class performance during the first week. (ibid)

2.2 Reliability

The second criterion of a good test is reliability. Charles Anderson states that a test is not valid if it is not reliable. (Charles Anderson: 1995)

It also refers to "how dependably or consistently a test measures a characteristic." (HR-Guide: 2018). In other words, a test measures reliability if it gives "similar scores for aperson who repeats the test …" (ibid). Suharsimi Arikunto states that a test is "… trusted if it gives … consistent result(s) when it is tested repeatedly." (Suharimi Arikunto: 1999)

2.3 Index Difficulty

The index difficulty describes a measure of "the *proportion* of examinees who answered the item correctly …". That's why in some contexts it is known as the "*p-value"*. The p-value might be called item easiness index because it measures the proportion of test takers who correctly answered the same test item (Professional Testing: 2020). This index can range between 0.0 as the lowest and 1.0 as the highest. So when the index difficulty is 1.0, this is an indication that a particular test item is too easy. (Anas Sudijono: 2007).

3. Distracters

Distracters have been described in details under the subtitle "Multiple-choice questions". However, it is important to know that the proportion of examinees who select each of the distractors can be very informative. When too many test takers select a distracter other than selecting the key answer, this distracter should be examined to decide whether or not it has been mis-keyed or double-keyed". On the other hand, when examinees fail to "select a given distractor", this may be an evidence that it is implausible or too easy. (Professional Testing: 2020)

4. Conclusion

It can be concluded throughout this research that tests in general is viewed as an effective way of determining the worth of a program at the end of instruction. Also, tests, test types, test functions, classification of tests based on form and method, multiple-choice tests as well as the criteria of a good tests all of them have been briefly discussed. There are other classifications of tests in based on their relationship with "item analysis". That is an agenda and a recommendation for more research projects. In this research, multiple-choice questions have been presented and

discussed due to their strong connection to objective tests. Hopefully, this research can provide good pieces of information for other researchers in this vital topic.

References

Anas, S. (2005). Pengantar Evaluasi Pendidikan. *Jakarta: PT. Raja Grafindo Persada.*

Anas, S. (2007). Pengantar Evaluasi Pendidikan. *Jakarta: Raja Grafindo.*

Charles, A. (1995). Caroline Clapham and Dianne Wall, *Language Test Construction and Evaluation*, British: Cambridge University Press.

David, P. H. (1969). Testing English as Second Language. *New York: Mc Graw Hill Book Company.*

Er, N. S. (2012). Perceptions of Turkish High School Mathematics Teachers Regarding the 2005 Curricular Changes and Their Effects on Mathematical Proficiency and University Entrance Exam Preparation. *ProQuest digital dissertation. Ohio University, Athens, Ohio.*

Gronlund, E. N. (1981). Measurement and Evaluation in Teaching (4th ed.). *New York: Macmillan Publishing Co., Inc.*

Gronlund, N. E., & Linn, R. L. (1990). Measurement and evaluation in teaching (6th ed.). *New York: Macmillan Publishing Company.*

Heaton, J. B. (1974). Writing English Language Test. *London: New Jersey.*

Heaton, J. B. (1998). *Writing English Language test*. Longman.

HR-Guide, (2018). Retrieved from https://hrguide.com/Testing_and_Assessment/Reliability_and_Validity.htm

M. Buchari in Suharsimi Arikunto, (1987). Dasar-dasar Evaluasi Pendidikan, (Jakarta: BinaAksara, 1987).

M. Buchari M. Ed, (1980).Tehnik-tehnik dalam Evaluasi Pendidikan, (Bandung, 1980).

McNamara, T. (2000). Language Testing: OXFORD UNIVERSITY PRESS:; Second Edition. Retrieved from https://books.google.iq/books?hl=ar&lr=&id=RuxUkltYl_UC&oi=fnd&pg=PR11&dq=What+is+testing+&ots=F4DfxTjpCx&sig=Mg_4HU7PbvYgamt8s6JVsiRxemE&redir_esc=y#v=onepage&q=What%20is%20testing&f=false

NgalimPurwanto, M. (2000). *Prinsip-Prinsi pdan Teknik Evaluasi Pengajaran* .Cet. 9. Bandung: Remaja Rosdakarya.

Popham, W. J. (1978). Criterion-referenced measurement. *Englewood Cliffs, NJ: Prentice-Hall.*

Popham, W. J. (2020). Test Better, Teach Better. Retrieved from http://www.ascd.org/publications/books/102088/chapters/The-Links-Between-Testing-and-Teaching.aspx

Professional Testing (2020). Retrieved from https://www.proftesting.com/test_topics/steps_9.php

Race, P. (2005). Making Learning Happen. *London: SAGE.*

Rahmadi, N. (2012). The Quality of Tests in the Textbook. *Palangka Raya: STAIN Palangka Raya.*

Rini, Y.(2006). Kemampuan Guru Menentukan Nilai Akhir Mata Pelajaran PAI di SMA Palangka Raya. *Skripsi, Palangka Raya: STAIN.*

Slameto, E. P. (2001). Cet.2. *Jakarta: Sinar Grafika Offset.*

Suharsimi, A. (1999). Dasar-Dasar Evaluasi Pendidikan. *Jakarta: Bumi Aksara.*

Thoha, C. (2003). Tehnik Evaluasi Pendidikan. *Jakarta: Raya Grafindo Persada.*

Wilmar, T. (1988). Evaluation of Students' Achievement, *Jakarta: Departemen Pendidikan dan Kebuyaan.*

Wilmar, T. (1998). Evaluation of students achievement. *Jakarta:Depdikbud.*

Woodford, E. P. (1980). Foreign Language Testing. *The Modern Language Journal, 64*. https://doi.org/10.1111/j.1540-4781.1980.tb05173.x

Copyrights