International Journal of Modern Language Teaching and Learning
Available online at www.ijmltl.com. *Vol. 1, Issue 3, 2016, pp.111-115*

Scoring in Testing: Review of Related Literature First

Sorour Zeigham
Department of ELT, Ayatollah Amoli Branch, Islamic Azad University, Amol, Mazandaran, Iran

Corresponding author’s email: sorour.zeigham68@gmail.com

ABSTRACT

The role testing in evaluating the performance of language learners is inevitable and crucial. The outcomes of language tests are usually reported as numbers or scores, and it is these scores, consequently, that tests users will make use of. In this regard, the role of scores is considerable in demonstrating language results. The scores of tests are applied to make a decision about a performance of a learner, the methods applied to get these score are a fundamental part of the measurement process. According to the significance and functions of test scores, the main purpose of the present study was to seek out the conceptual framework of scores in language testing.

KEYWORDS:*Scoring in test*, *language teaching*, *language testing,* learner behavior

INTRODUCTION

Language learning is a crucial process in individuals' life. The main importance of this process is to have a power to communicate. This assists one's to achieve his/her goals in different fields like politics, culture, business, and so on. For assessing the performance and progression of the learner, it is necessary record their performance by tests. It is essential to present a definition firstly. A test is a tool or systematic procedure for observing and describing one or more features of a learner, applying either a numerical scale or categorization scheme.There are different types of tests in the language testing process such as placement, proficiency, diagnostic tests and so on and each of these tests are utilized for various purposes. Some of these purposes are such as recognizing the weaknesses and strength of the learners, determining the current level of the learners, the ability to acquire a skill and so forth. Hence, tests has a special place in language teaching and learning.

One important part of language testing is scoring. The outcomes of language tests are usually reported as numbers or scores, and it is these scores, consequently, that tests users will make use of. According to Anastasi (1990), score or test score is a number or other quantitative (numerical) value applied to represent: a) a person response, where the term response is meant to limit any measurable act of a person; and b) a total or summed value based on a number of a person's score. The scores of tests are applied to make a decision about a performance of a learner, the methods applied to get these score are a fundamental part of the measurement process.

According to the significance and functions of test scores, the main purpose of the present study is to seek out the conceptual framework of scores in language testing. This process, which plays a vital role in insuring that the scores of the test are reliable and that the uses made of them are valid, includes two stages: a) defining the construct theoretically; and b) defining the construct operationally.

DEFINITIONS OF SCORING

According to Lyman (1963), scoring or scorability means that each item in the test has its own mark related to the distribution of marks given by. In specifying what scoring method to apply, it is needful to consider two dimensions of development of test: the theoretical definition of the construct to be measured, which is a section of the design statement, and the task seek specifications, which are the blueprint section (Bachman & Palmer, 1996). The construct definition of the scoring specifies the type of score to be reported, whether this is a profile of scores of various realms of language ability, a single composite score, or both (Bachman & Palmer, 1996).

Based on the different types of scoring methods, in one approach, the score is defined as the number of test tasks successfully completed, in order that the number of correct responses are added up.

THE NATURE OF SCORING

Scores are important in the process of testing. The scores on a testing means are expected to be a reflection of the construct we plan to measure, for instance language proficiency, and nothing else. Although, in practice, furthermore other factors affect test performance. For instance, if a candidate's writing proficiency is tested by means of a writing prompt, eliciting a written composition, the writing score the candidate receives will be the result of the candidate's writing proficiency, the features of the writing prompt, the rating criteria, the rater's severity, the interactions between these factors and probably some random other factors as well (Fulcher& Davidson, 2012).

Scoring is related to the *how much* or *how good* of language testing. It can be said that draw inferences from performance on a task to the ability is executed to control turn taking in social interaction, and the tasks are designed to elicit evidence, but this does not tell us how much evidence is necessary, or how much of the ability is present.

STEPS OF SCORING A TEST

According to Gronlund and Linn (1990), scoring a test consists of three steps:

a) Defining the construct theoretically: In defining the construct the test developers requires to make a conscious and deliberate option to determine specific elements of the ability or abilities to be measured in a way that is proper to a specific testing situation. Particular definitions of the abilities, or constructs are required for three objectives:

. To provide a foundation for using test scores for their intended objectives

. To guide test construction attempts

. To enable the test developer and user to demonstrate the construct validity of these interpretations (Bachman, 1990)

Therefore, the test developer requires to decide what abilities to involve and not include in the construct definition. The constructing way for a specific testing situation will specify which areas of language ability we require to score and the role, if any, of topical knowledge in the scoring method (Bachman, 1990).

b) Defining the construct operationally: an operational definition of different construct is frequently involved in this phase. It means that the test developer and task writers have to be as obvious as possible as to what a given test task is assumed to be measuring (Anastasi, 1990). This will be effective for test users in interpreting the test outcomes. Constructs are operationally defined to specify the kind of planned response, and this has clear implications for scoring. Therefore, tasks intended to elicit a selected response can generally be scored of the language of the intended response (for instance length, organizational features, and pragmatic attributes) in specifying the particular scoring method to be applied.

c) Establishing a method of quantifying responses to test tasks: Quantification is the assigning of numbers to individuals' responses. In determining the scoring method (the criteria by which test takers' responses are assessed and the procedures followed to arrive at a score). It is essential to specify how to quantify test takers' responses (Anastasi, 1990).

Therefore, scoring the necessary phase to arrive at a measure, furthermore any quantitative, descriptive information obtained from the test takers' responses (Bachman & Palmer, 1996).As with most decisions in designing and developing language tests, the initial decisions about scoring have to be checked by actually giving the test and evaluating the testing procedures and results in terms of their usefulness. It means that the initial decisions about scoring such as the test, observed test takers, scored their responses, and analyzed and interpreted these outcomes have to be taken into account tentatively. Therefore, in the development of scoring method, a critical phase is to try out the test tasks with one or more groups of individuals who are representative of the intended test takers, score their responses, and analyze the outcomes.

DIFFERENT WAYS OF SCORING ITEM TYPE

There are different ways of scoring item type. According to Alderson, Clapham and Wall (1995), four different ways of scoring are presented for item type:

. *Exact match:* The first type is exact match. According to this type, permitting one mark for each sentence in the correct place in the sequence.

This is the clearest way to score the item, but one that suffers from the problems we have already outlined above.

. *Classic:* The second type of scoring item is classic type. In this type of scoring, one mark for an exact match, one mark if the previous sentence is the correct one in the sequence, one mark if the following sentence is the next in the sequence, andone mark if the sentence is the last in the sequence and no sentence follows it (an 'edge' score).

. *Added value:* The third type of scoring item is called added value. This type of scoring item is as same as the classic, but with no score for an exact match.

. *Full pair:* The last type of scoring type is full pair. In this type of scoring, the sum of previous and next scores, excluding an 'edge' score are regarded.

According to Fulcher (2010), the first type, named exact match has an advantage and its advantage is that this type is simple and easy to score and calculate. The classic procedure keeps the exact match component, but gives scores for correct pairs and triplets in the right order, even if they are not in the same position. The added value method simply removes credit for the correct position, and gives credit for correct sequencing in pairs or triplets. The full pair method also credits pairs and triplets, but has no exact match score and no 'edge' score.

SCORING AS THE NUMBER OF TASKS SUCCESSFULLY COMPLETED

GENERAL CONSIDERATIONS

Tasks that include items can be utilized to measure particular language knowledge areas, in addition to the ability to apply language in receptive language use tasks, listening and reading comprehension. In tests that comprise tests, typically test takers needed either to choose an answer from among several options (selected response) or to produce a restricted language sample in response to the task (limited production response). For both of answering types, the most regularly applied approach to scoring is to add up tasks number successfully completed, the number of correct responses. Assuming that the task is adequately well defined by the way the item is designed and written, the basic considerations for scoring are:

Determining the criteria for what forms a correct response to the task, andSpecifying ways for scoring the responses, that is, making decisions whether responses will be scored as right or wrong or in terms of degrees of correctness (partial credit scoring) (Bachman & Palmer, 1996).

DETERMINING THE CRITERIA FOR CORRECTNESS

Different criteria for correctness can be applied with both chosen and restricted production responses, relying on language knowledge areas to be evaluated. In one tradition that still informs a lot of languages tests today, it is regarded necessary to apply one criterion for correctness, in the effort to attain a 'pure' measure of a particular language knowledge areas (Bachman & Palmer, 1996). Therefore, for a specified item to assess merely grammatical knowledge, for instance, one might logically apply grammatical accuracy as the single criterion for correctness. This can be carried out that quite simply with chosen response items, by supplying merely one alternative that is grammatically correct, the 'key', while all the other options, or 'distractors', are grammatically incorrect.

While applying one criterion may work logically well with chosen responses, this may provide problems with limited production responses (Bachman & Palmer, 1996). For instance, applying grammatical accuracy as a single criterion for correctness may result in counting as correct some answers that are lexically improper.

In measuring the items about lexical knowledge, the test developer might not regard grammatical accuracy at all, but use meaningfulness as the single criterion.

SPECIFYING PROCEDURES FOR SCORING THE RESPONSES

As Bachman (1990) states that both chosen and limited production responses can be scored in one of two items: right/wrong or partial credit. With right/wrong scoring, a response receives a score of '0' if it is wrong and '1' if it is correct. With partial credit scoring, responses can be scored on several levels, ranging from no credit '0' to full credit, with different partial credit levels in between. Most language testers have tended to favor right/wrong scoring for chosen and limited production responses, widely because the techniques accessible for statistically analyzing the test items features: difficulty and discrimination were designed to work best if responses were scored due to a single criterion for correctness.

MULTIPLE RIGHT/WRONG SCORES FOR RESPONSES TO A SINGLE ITEM

According to Bachman and Palmer (1996), with partial credit scoring we utilize multiple criteria for correctness to arrive at a single score for each item: we give full credit for a response that satisfies all criteria for correctness, partial credit for responses that satisfy some of the criteria, and no credit for answers that satisfy none of the criteria. Multiple scores and partial credit scoring include two merits: a) they offer the test user the potential for taking more information about responses, and therefore more information about the areas of test takers of merits and demerits, than does giving a single right/wrong score. With the multiple-score approach, it is probable to note separate scores for various realms of language ability that are tested. Reporting separate scores may be specifically effective where tests are to be applied for diagnostic objectives, that is, for providing feedback to learners on their realms of merits and demerits and to teachers on which realms of the syllabus seem to be working efficiently to promote learning and which realms require enhancement (Bachman & Palmer, 1996).

SCORING AS LEVELS OF LANGUAGE ABILITY

As Bachman and Palmer (1996) point out that there are different types of scales in language ability: a) global scales of language ability; b) Analytic scales of language ability; c) criterion-referenced scales of language ability; d) Primary trait scales and e) Multiple trait scoring.

GLOBAL/HOLISTIC SCALES OF LANGUAGE ABILITY

One traditional approach to develop rating scales of language proficiency is related to the view that language ability is considered as a single unitary ability, and yields a single score, named a 'global' rating. In other words, in this type of scale, a single score is awarded, which reflects the overall quality of the performance. Although, most of these scales include multiple 'hidden' elements of language ability. Generally, holistic scales are fairly easy to apply and with extensive training high levels of inter-rater reliability can be achieved.

According to Bachman and Palmer (1996), global scales encounter some problems:
a) Inference problems,
b) Difficulties in assigning levels, and
c) Differential weighting of elements.

THE INFERENCE PROBLEM

Using global scales makes them difficult to know what a score reflects: multiple realms of language knowledge, topical knowledge, or multiple language use situations.The second problem with global scales, associated with the inference problem, is that raters applying them frequently have difficulty in assigning levels.

CRITERION-REFERENCED SCALES OF LANGUAGE ABILITY

With global scales, always there is the probability that different raters (or the same rater on different situations) may either consciously or unconsciously weigh the hidden elements variously in arriving at their single rating.

ANALYTICAL SCALES OF LANGUAGE ABILITY

This approach is related to two tenets:
a) The operational definitions in the scales are associated with theoretical definitions of the construct. These may be either theory-based or syllabus-based componential definitions of language ability.
b) Secondly, the scale levels are referenced to determined levels in diverse realms of language ability.

The use of analytic scales include two practical merits: first it permits us to provide a 'profile' of the language ability realms that are rated. A second merit is that analytic scales tend to show what raters actually carry out when rating samples of language use.

CRITERION-REFERENCED SCALES OF LANGUAGE ABILITY

Criterion-referenced scales of language ability is based on two tenets. The principal merit of criterion-referenced scales is that they permit us to make inferences about how much language ability a test-taker has, and not merely how well the test takers performs relative to other individuals, involving native speakers. The second tenet is to define scales practically in terms of criterion ability levels (Bachman & Palmer, 1996).

PRIMARY TRAIT SCALES

In this type of scale, there is a single score, but the descriptors are enhanced for an individual prompt (or question) that is applied in the test. In this regard, every prompt is enhanced to take a particular response, perhaps an argumentative essay in an academic situation, for instance. The rating of this scale indicates that the particular qualities specified in writing samples at a number of levels on the scale. At each level, writing Samples a are supplied to exemplify what is targeted by the descriptor. This is an improvement over holistic scales, but a different scale has to be developed for each prompt or task type, which increases the investment of time and resources in scale development (Fulcher, 2010).

MULTIPLE TRAIT SCORING

Multiple trait scoring, unlike other types of scales, needs raters to award two or more scores for diverse traits or characteristics of the speech or writing sample. Regularly, the traits are prompt-type specific, as in primary trait scoring. The argument in favor of this kind of scoring is that richer information is supplied about each performance. In the case of an essay this may involve traits like as organization, coherence, cohesion, content, and so on (Fulcher, 2010).

CONCLUSION

A test is a tool or systematic procedure for observing and describing one or more features of a learner, applying either a numerical scale or categorization schemein the process of language teaching and learning. The outcomes of language tests are usually reported as numbers or scores, and it is these scores, consequently, that tests users will make use of. The scores of tests are applied to make a decision about a performance of a learner, the methods applied to get these score are a fundamental part of the measurement process. According to the significance and functions of test scores, the main purpose of the present study was to seek out the conceptual framework of scores in language testing. Four different ways of scoring are presented for item type: a) exact match; b) classic; c) added value; and d) full pair. Different criteria for correctness can be applied with both chosen and restricted production responses, relying on language knowledge areas to be evaluated. As Bachman and Palmer (1996) point out that there are different types of scales in language ability: a) global scales of language ability; b) Analytic scales of language ability; c) criterion-referenced scales of language ability; d) primary trait scales and e) Multiple trait scoring.

Conflict of interest
The authors declare no conflict of interest.

REFERENCES

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation.*New York: Cambridge University Press.

Anastasi, A. (1990). *Psychological Testing*. New York, Macmillan Publishing Company.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford Oxford University Press.

Bachman, L. F., & Palmer, A. (1996). *Language testing in practice: Designing anddeveloping useful language tests*. Oxford: OUP.

Fulcher, G. (2010). *Practical Language Testing*. London: Hodder Education.

Fulcher, G., & Davidson, F. (2012). *The Routledge handbook of language testing*.Routledge.

Gronlund, N. E., & Linn, R. L. (1990). *Measurement and evaluation in teaching* (6thed.).New York: MacmillanPublishing Company.