CONSTRUCTING A STANDARDIZED TEST[1]

Sofendi, M.A., Ph.D.[2]

This paper is aimed at describing what a standardized test is and how it is constructed. A standardized test should yield good validity and reliability as well as practicality. This kind of the test is considered as a good instrument of measurement. Constructing a standardized test requires some steps. These steps lead to getting good validity, realibility, and practicality of the test. The only valid and reliable test can be an interpretable mesurement.

Key words: Test, Validity, Realibility, Practicality

I. Introduction

Testing has been side by side with teachers. This implies that, one of them, teachers use tests (teacher-tailored tests) in almost every occasion in their academic routines. They may test their students regularly, for example, monthly, termly and/or annually. However, there are at least two fundamental questions arise in line with the teacher-tailored tests, that is "Are the tests valid?" and "Are the tests reliable?". Most teachers do not really care about reliability of their tests, but they implicitly care about validity of their tests. This is because they usually devise tests on the basis of what they have taught without thinking the consistency or stability of the test. Therefore, they very often do not feel confident if they have to answer some fundamental questions related to the ways they rank their students. This is mainly because their tests do not have any information on reliability.

Validity and reliability of tests are fundamentally important. This is because if the tests are not valid and reliable, then the results of the tests are not interpretable. This means that the tests should have validity and reliability in order that the results can be used as interpretable measurements.

This paper is generally concerned with a test which has good validity and reliability. This kind of test is generally known as a standardized test. This is because a standardized test is, among others, a valid and reliable test. The focus of this paper will be on constructing a valid and reliable test. Therefore, this paper will first of all begin with a short description on what a standardized test is, and then it will describe the ways that should be done in line with getting validity and reliability of a standardized test, and finally it ends with some conclusions.

---

[1] This paper is presented in Semirata BKS-PTN at Swarna Dwipa Hotel, Palembang on 22 and 23 July 2009

[2] Lecturer at the English Education Study Programme, Language and Arts Education Department, Faculty of Teacher Training and Education, University of Sriwijaya and Director of Sriwijaya University Language Institute

II. A Standardized Test

The term “test” is a simple and widely used but it paradoxically somewhat vague. Ahmann and Glock (1981) generally define a test so broadly as to include some evaluation procedures that yield only verbal descriptions of student traits, and specifically, it is nothing more than a group of questions to be answered or tasks to be performed. Unlike Ahmann and Glock, Cronbach (1970) provides a bit more specific definition. They claim that a test refers to any systematic procedures for observing procedures for observing a person’s behaviour and describing it by means of a numerical scale or a category system (Cronbach, 1970). Furthermore, Brown (1987) defines a test even more specific than those described earlier. He defines a test as a method of measuring a person’s ability or knowledge in a given area. All these three definitions at least provide one idea that a test is an instrument that can be used for the purpose of measurement.

A standardized test is generally considered as a good test. A good test, according to Harris (1969) and Brown (1987), has to have three characteristics – reliability, validity and practicality. In line with three characteristics, Harris (1969) claims that reliability refers to stability of test scores, validity concerns with the questions “What precisely does the test measure?” and “How well does the test measure?”, and practicality relates to economy (cheap), ease of administration and scoring, and ease of interpretation. However, Brown (1987) provides information on these three characteristics a little bit different from those claimed by Harris. He says that a reliable test is a test that is consistent and dependable, validity is the degree to which the test actually measures what it is intended to measure, and a test ought to be practical within the means of financial limitations, time constraints, ease of administration, and scoring and interpretation.

Among those three characteristics, practicality is considered not really a fundamental pre-requisite. This is because if a test, for example, is expensive, difficult to administer, score and interpret does not affect the reliability and validity the test itself. Therefore, practicality may, to some extent, be neglected, but reliability and validity of test in any condition cannot be avoided. This is because a test as an instrument has to be valid and reliable in order that the results can be interpretable. Therefore, Adkins (1974), Kline (1975), Ahmann and Glock (1981), Gronlund (1988), and Rust and Golombok (1989) claim that a good test must be reliable and valid.

In line with reliability and validity, Rust and Golombok (1989), and Hieronymus, Lindquist and France (1988) provide clearer definitions than Harris (1969) and Brown (1987) do. Validity is concerned with whether the test is measuring what is supposed to measure (Rust and Golombok, 1989; and Hieronymus, Lindquist and France,1988). Reliability is concerned with the extent to which test scores measure “true” variance and is expressed

numerically in the form of a reliability coefficient ranging from 0 – 1 (Hieronymus, Lindquist and France,1988).

In short, a standardized test as a good test should be valid, reliable and practical. However, practicality may possible be neglected but not validity and reliability. In other words, a standardized test must at least be valid and reliable.

III. Constructing A Standardized Test

Constructing a standardized test is equivalent to constructing a good test. Constructing a good test should be done scientifically reasonable and widely acceptable. Besides, a good test must have distinct features that make it different from bad one. It must be reliable and valid. Therefore, to determine the merit of any test, Downie and Health (1974) claims that test results must be subjected to an item analysis. The analysis of test item, as Downie and Health further claims, leads to three kinds of information: (1) difficulty of the item (p) – proportion of individuals who answer an item correctly, (2) the discrimination index of the item (r) – a measure of how well the item separates two groups (good and poor ones), and (3) the effectiveness of the distracters – how the incorrect responses in the multiple-choice item are working. The results of analysis of test item finally provide information of reliability and validity of the test.

But, before describing how to get a reliable and valid test, first of all, general ways of constructing a test will be described. In constructing tests, there may be slightly different steps that have to be taken. This is because different tests may require different prerequisites, e.g. constructing a norm-referenced test may have different steps from constructing a criterion-referenced test.

In general, among others, Harris (1969), Walsh and Bezt (1995) and Sofendi (1998) suggest general ways of constructing a good test.

Harris (1969) proposes seven general steps in constructing a test. The steps are (1) planning the test, (2) preparing the test items and directions, (3) reviewing the items, (4) pretesting the materials, (5) analyzing the pretest results, (6) assembling the final form, and (7) reproducing the test.

Unlike Harris, Walsh and Bezt (1995) only claims six general steps in constructing a test. They are: (1) beginning with a careful, detailed definition of the attribute, construct or characteristics to be measured, (2) developing test items that are related to the content (i.e. definition), (3) administering the items to a preliminary sample of subjects – the subjects in this group should be representative of the population of subjects for whom the test itself is intended, (4) refining the items, refining the items means eliminating items

that do not have the properties we had hoped for and selecting items that have particularly desirable properties, through item analysis (to find the item difficulty and item discrimination) and expert judgment (to get information on the appropriateness of test item(s), (5) administering the revised test to a new sample of subjects, and (6) examining the evidence for reliability and validity, and compute normative data.

However, sofendi (1998) claims more steps than those proposed by the two earlier experts. He suggests ten general steps that should be done in line with constructing a test. The steps are as follows: (1) identifying and classifying objectives and areas, (2) selecting and determining the test type, (3) determining the total number of test items and test length, (4) deciding the levels of cognitive domains and weighing the test items, (5) devising the test items, (6) asking for experts' judgements on appropriateness and difficulty levels of test items, (7) revising the test items, (8) trying out the test, (9) analysing the results, and (10) producing the final test.

The above three general steps proposed by Harris (1969), Walsh and Bezt (1995), and Sofendi (1998) can be summarised into five very general steps. They are preparing, devising, trying out, analysing and producing the test. All these steps are ultimately aimed at finding out the validity and reliability of the test. For example, preparing and devising the test are aimed at getting a test draft. This test draft is then tried out to get some data that will be used to find out the validity and reliability of the test in the analysing step. If good validity and a good reliability coefficient of the test have been obtained then the final form of the test can be produced.

As this paper focuses on the validity and reliability of standardized test, therefore, the ways of getting the validity and reliability will be explored further.

Validity of test items can be obtained by asking for experts' judgements. Experts' judgments can be obtained before the test is tried out. However, before asking for experts' judgments, the test maker should first of all devise ranks/classifications of appropriateness and difficulty levels of test item so that the experts can give reasonable judgements on the test items. The validity of the test items may be different from one kind of test to another. However, in general, according to Rust and Golombok (1989) there are five kinds of validity can be obtained from one test. They are (1) face validity - it concerns the acceptability of the test items, to both test user and subject, for the operation being carried out, (2) content validity - it examines the extent to which the test specification under which the test was constructed reflects the particular purpose for which the test is being developed, (3) predictive validity - it is the major form of statistical validity and is used wherever tests are used to make predictions, (4) concurrent validity - it is statistical in conception and describes the correlation of a new test with existing tests which purport

to measure the same construct, and (5) construct validity - it is the primary form of validation underlying the trait related approach to psychometrics.

Reliability of a test, according to Rust and Golombok (1989) can be obtained through one of four techniques. The four techniques are (1) Test-Retest Reliability - it involves administering the test twice to the same group of respondents, with an interval between the two administration of, say, one week. This would yield two measures for each person, the score on the first occasion and the score on the second occasion. A Person product-moment correlation coefficient calculated on these data would give us a reliability coefficient, (2) Parallel Forms Reliability - two versions of a test are linked in a systematic manner and are intended to measure the same construct (e.g. 2 + 7 in the first version of an Arithmetic test, and 3 + 6 in the second). Two tests constructed in this way are said to be parallel. To obtained the parallel forms reliability, each person is given both version of the test to complete, and we obtain the reliability by calculating the Pearson product-moment correlation coefficient between the scores for the two forms, (3) Split Half Reliability - it involves administering the test once. Then, each paper (a test) is randomly split in two, e.g. odd-numbered items and even-numbered items or other splits, to make half-size versions of the test. So, for each individual two scores are obtained, one for each half of the test, and these are correlated with each other, again using the Pearson product-moment correlation coefficient to get the reliability of half of the test. To obtain the whole test, we apply the Spearman-Brown formula, and (4) Inter-Rater Reliability or Inter-Marker Reliability - when different markers of the same essay tend to give different marks, or different interviewers may make different ratings of the same interviewee within a structure sets of ratings respectively using the Person product-moment correlation coefficient between the scores of the two raters.

In line with a standardized test as experts claim as a good test, the test must be valid and reliable. The test can generally be considered valid and reliable if it contains at least four out of five types of validity (face validity, content validity, predictive validity and construct validity), and a reliability coefficient from 0.90 to 1.

The following is an example of how to get the realiability coefficient of the test:

Calculating Reliability Coefficient of a Test by Using a Split-half Method

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
X	Y	x	y	Zx	Zy	ZxZy	x^2	y^2	xy
20	12	7	2	1.61	0.54	0.8689	49	4	14
18	16	5	6	1.15	1.62	1.8637	25	36	30
16	10	3	0	0.69	0.00	0.0000	9	0	0
15	14	2	4	0.46	1.08	0.4968	4	16	8
14	12	1	2	0.23	0.54	0.1242	1	4	2
12	10	-1	0	-0.23	0.00	0.0000	1	0	0
12	9	-1	-1	-0.23	-0.27	0.0621	1	1	1
10	8	-3	-2	-0.69	-0.54	0.3726	9	4	6
8	7	-5	-3	-1.15	-0.81	0.9315	25	9	15
5	2	-8	-8	-1.84	-2.16	3.9744	64	64	64
∑X=130	∑Y=100	Sx=4.34	Sy=3.71			∑ZxZy=8.6947	$\sum x^2$=188	$\sum y^2$ =138	∑xy = 140
X=13	Y=10								

Notes:

- Column 1 : odd-numbered items
- Column 2 : even-numbered items
- Column 3 : Deviation of each of X scores from the mean (e.g. 20 -13 = 7)
- Column 4 : deviation of each of Y scores from the mean (e.g. 12 – 10 = 2)
- Column 5 : standard deviation of x (column 1), (e.g. 7 : 4.34 = 1.61)
- Column 6 : standard deviation of y (column 2), (e.g. 2 : 3.71 = 1.54)
- Column 7 : product of the two Z scores (e.g. 1.61 x 0.54 = 0.8694)
- Sx (4.34) : the standard deviation of x or the standard deviation of column 3
- Sy (3.71) : the standard deviation of y or the standard deviation of column 4

- The Person Product-moment Correlation Coefficient : $r = \frac{\sum ZxZy}{N}$

- The Spearman-Brown Formula : $r_{tt} = \frac{2\ roe}{1 + roe}$

The scores of ten individuals on two variables, X and Y (columns 1 and 2). Beneath each of these is the mean. In column 3 is the deviation of each of the X scores from the mean of X, and in column 4 we find the deviation of each of the Y scores from the mean of Y. Beneath these two columns are the standard deviations of the columns. In columns 5 and 6 are the standard scores for each of the scores in columns 1 and 2. These were obtained by dividing each score in column 3 by Sz (4.34) and each value in column 4 by Sy (3.71). Here we are using the usual formula for Z. Z = x/s. Column 7 is the product of the two Z scores, the product of the values in columns 5 and 6. These are summed, and the Pearson r is obtained by the following formula, which describes r as the mean Z score product:

$$r = \frac{\sum ZxZy}{N}$$

By substituting into this formula, we obtain:

$$r = \frac{8.6947}{10} = 0.869 \text{ or } 0.87$$

$$r_{tt} = \frac{2\ (0.869)}{1 + 0.869} = 0.93 \text{ (this is the reliability coefficient of the test)}$$

IV. Conclusions

Having described briefly what a standardized test is and how to construct a standardized test, two conclusions can be drawn, that is (1) a standardized test as a good test must be valid, reliable and practical but practicality, to some extent, can be neglected, and (2) validity and reliability of standardized test can only be obtained through reasonable steps in its construction.

V. References

Adkins, D.C. (1974) Test Construction: Development and Interpretation of Achievement Test. Columbus: Bell and Howell Company.

Ahmann, J.S. and Glock, M.D. (1981) Evaluating Student Progress: Principles of Tests and Measurements. London: Allyn and Bacon, Inc.

Brown, H.D. (1987) Principles of Language Learning and Teaching. New Jersey: Prentice-Hall, Inc.

Cronbach, L.J. (1970) Essentials of Psychological Test. New York: Harper Routledge.

Downie, N.M. and Heath, R.W. (1974) Basic Statistical Method. London: Harper and Row Publishers.

Gronlund, N.E. (1988) How to Construct Achievement Test. New Jersey: Prentice-Hall, Inc.

Harris, D.P. (1969) Testing English as a Second Language. New York: McGraw-Hill Book Company.

Hieronymus, A.N., Lindquist, and France (1998) Richmond Test of Basic Skills: Administration Manual. Berkshire: The NFER-NELSON Publishing Company Ltd.

Kline, P. (1975) Psychological Testing: The Measurement of Intelligence, Ability and Personality. London: Malaby Press.

Rust, J. and Golombok, S. (1989) Modern Psychometrics: The Science of Psychological Assessment. London: Routledge.

Sofendi (1998) The Effects of Groupwork on Mathematics Attainment in Indonesian Primary Schools (Ph.D. Thesis). London: University of London.

Walsh, W.B. and Betz, N.E. (1995) Tests and Measurement. New Jersey: Prentice Hall.