World Transactions on Engineering and Technology Education
Vol.4, No.2, 2005

Assessment of learning outcomes: validity and reliability of classroom tests

Maizam Alias

Kolej Universiti Teknologi Tun Hussein Onn
Johor Darul Takzim, Malaysia

ABSTRACT: Teachers in engineering routinely design and administer classroom tests to their students for decision-making purposes. To be of real value in decision-making, these tests must be valid and reliable. Test validity and reliability may be achieved by taking a systematic approach to test design. In this article, the author proposes and discusses measures that teachers could take in order to help them enhance the validity and reliability of their classroom tests, taking examples from the teaching and learning of structural design in civil engineering. A sample spreadsheet in *Excel* is provided that may be used by teachers to get a quick estimate of the reliability of their classroom tests.

INTRODUCTION

Assessment entails the systematic gathering of evidence to judge a student's demonstration of learning. Teachers can then judge whether a student has learned what they are expected to learn by securing valid and reliable information through various assessment methods. The assessment method chosen would depend on the learning domain that is of interest, which could be the cognitive, affective or psychomotor domains [1]. Examples of learning in the three domains are given in Table 1.

Table 1: Examples of learning in the cognitive, affective and psychomotor domains [2].

Domain	Learning Outcomes
Cognitive	Able to solve simultaneous equations
Affective	Choosing to learn from own and other peoples' experiences by ensuring similar mistakes are not repeated, and incorporating past successes into current design where appropriate
Psychomotor	Manoeuvring a computer mouse to produce the desired effect on the computer screen when using a Computer Aided Design package for drawing

Assessment tools that can be used to assess learning include an achievement test for the cognitive domain, an attitude questionnaire for the affective domain and a checklist for the psychomotor domain. In this article, the author focuses on the assessment of learning in the cognitive domain since this is the most frequently assessed domain for classroom learning.

For the purpose of this article, the term *classroom test* will be used in place of achievement test to emphasise its classroom application. A classroom test is defined as any set of questions that is specifically designed by teachers to measure an identified learned capability (or set of learned capabilities) and administered by teachers to their students in classroom setting.

Classroom tests are routinely designed and administered by teachers to assess students' learned capabilities, and output from classroom tests are often used to support decision-making, such as in giving grades to students or assigning students to remedial classes. In order to be of real value in decision-making, these classroom tests must possess two important characteristics, namely: validity and reliability. A discussion of some of the issues that teachers need to look into, plus some of the practical measures that teachers can take to enhance the validity and reliability of their classroom tests, is presented in this article.

VALIDITY AND RELIABILITY

Validity and reliability are two quality indicators for classroom tests. Validity refers to the degree to which a test is measuring what it is supposed to measure, while reliability is an indication of the consistency between two measures of the same test [3]. A test may be highly reliable but not necessarily valid, but a highly valid test is usually reliable.

Types of Validity

There are two types of validity that are most relevant to classroom tests, namely: face validity and content validity [3]. Face validity refers to the appearance of a test that looks like it is measuring what it is supposed to measure. Face validity is essential in ensuring that test-takers persevere and try their best on a test. A test that appears to be other than what it claims to be measuring – without face validity – may dissuade students from persevering with the test. Therefore, ascertaining whether a test possesses face validity does not require the opinion of an expert.

In contrast to face validity, a claim of content validity requires affirmation from an expert. The expert should look into whether the test content is representative of the skills that are supposed to be measured. This involves looking into the consistency between the syllabus content, the test objective and the test contents. If the test contents cover the test objectives, which in turn are representative of the syllabus, it could be said that the test possesses content validity.

For example, an English test paper is definitely not a valid instrument for measuring mathematical skills. An algebra test, on the other hand, is to a certain degree a valid measuring tool for mathematical skills because the ability to do algebra is an indicator of a person's mathematical skills. Still, the algebra test is not highly valid because mathematical skills are not confined to the ability to solve algebra problems alone. Therefore, to make the test paper highly valid, other indicators of mathematical skills must be included in the test paper. If the test is valid and reliable, a student who shows good mathematic skills on that particular test should also do equally well on other mathematical tests of similar content and objective. In other words, students do not just possess skills to solve the mathematical items that are given in the specific test.

To summarise, the decision on what to include in a test paper will depend on what the content of the syllabus is, as well as what the test objectives are. It is of utmost importance for teachers to appreciate that the degree of test validity depends on the test's coverage of the necessary objectives, which, in turn, depends upon the syllabus.

Types of Reliability

There are three types of reliability that are most relevant to classroom tests, namely: internal consistency, inter-scorer and intra-scorer reliability [3]. Internal consistency refers to the consistency of objectives among the items of a test. For example, consider a 10-item mathematical test that is supposed to measure students' ability to solve two variable algebra problems. In this case, the question of internal consistency refers to the answer to the question: are the 10 items measuring the same skill (ie students' ability to solve two variable algebra problems), or are the different items measuring something else entirely or others besides the stated objective?

Inter-scorer reliability refers to the consistency between the marks given by different teachers. Doubts upon inter-scorer reliability could arise when the same quality of answers is given different scores by different teachers. On the other hand, intra-scorer reliability refers to marks given by the same teacher on different occasions. An example of intra-scorer reliability at stake is when a teacher gets tired of marking and starts to give lower marks as time goes on. Consistent grading is essential in order to ensure the reliability of test scores.

Scorer reliability can be improved by a marking scheme or a scoring rubric that is prepared in advance and used to assist teachers in scoring answer scripts.

So what can be done to develop a valid and *reliable* test?

In order to achieve a certain degree of validity and reliability, the assessment and evaluation process has to be looked at in its totality, and the factors that may affect validity and reliability need to be identified. Typical activities in the classroom assessment and evaluation process are as follows:

- Deciding on a test's objectives;
- Designing and developing a test;
- Evaluating the test;
- Administrating the test.

At each stage, something could be carried out to enhance the validity and reliability of a test. The discussion below is based on these activities, starting with the decision on test objectives.

Deciding on a Test's Objectives

Determining a test's objective(s) is the first step in a test's construction process. The test objective is the criterion that will be used in order to judge whether a test is sufficiently valid or not. This objective is general in nature, which can be represented by a set of more specific objectives or item objectives to be identified through an analysis of the syllabus. Three examples of test objectives are as follows:

- Measure final year students' ability to solve calculus problems;
- Measure first year students' understanding of concepts and procedures in circuit design;
- Measure civil engineering students' ability to solve structural design problems that demand spatial visualisation abilities.

The key phrases, *calculus problems*, *concepts and procedures in circuit design* and *structural design problems that demand spatial visualisation abilities* determine the scope/content, item format and length of the test. If teachers as test designers are not clear of their objectives, they may end up measuring something other than what they wish to measure, that is, having an instrument that lacks validity.

Designing and Developing a Test

Designing a test is indeed a complex task. Many questions need to be asked and a lot of decisions need to be made at a number of stages along the way so as to increase the chances of meeting the criteria of a good test. In other words, the design and development stage of a classroom test holds the most possibilities for ensuring test validity and reliability. One of the most important steps in designing a test is constructing a table of specifications.

Constructing a Table of Specifications

As mentioned earlier, validity is concerned with how good a match is between what a test is supposed to measure and what it actually measures. Adequate content coverage is an important element of content validity. Constructing a table of specifications is one of the practical means towards achieving this objective.

A table of specification is a two-way table with the cognitive emphasis on the first row and contents in the first column. It can be constructed using a two level analysis described below.

The first level of analysis covers the following:

- Construct a two-way table with a list of topics in the first column and a list of cognitive emphases in the first row;
- Identify the topics/sub-topics and the corresponding cognitive emphasis to be tested;
- Estimate the percentage allocation for each topic.

The second level of analysis incorporates the following:

- Choose the appropriate item format (multiple choice (MC)/structured question (SQ)/long question or essay (LQ), etc) for the specific objective;
- Determine the number of questions for each specific objective;
- Check that the marks for each topic match the total weightage allocated.

An example of a table of specifications is given in Table 2 (taken from Alias [5]). This table of specifications forms the basis for designing a one-hour test on factual, concept and procedural knowledge of a beam, column and slab in structural design. The test covers the six cognitive skills as identified in Bloom's taxonomy [4]. By constructing a table of specifications, teachers are forced to consider in a systematic manner the learning objectives that need to be covered by their tests. Therefore, a test that possesses content validity is ensured.

Deciding on Item Format

The choice of item format depends upon several factors, with the item objective being the most important. Apart from the item objective, ease of scoring, ease of administration and the content coverage are also relevant factors in deciding on the item format. Common item format includes multiple choice, essay, structured and true/false. Certain formats are more suitable than others in meeting the item objective. For example, an essay question allows a student answering the question to demonstrate his/her depth of knowledge. On the other hand, essay questions are relatively more time consuming to mark and need greater efforts to ensure inter-scorer and intra-scorer reliability. In brief, when designing test items, a teacher has to balance the needs of the test objectives while also considering other practical constraints that may contribute to lower (or enhance) the test's validity and reliability.

Constructing Items

Once the format is chosen, the teacher has to construct the test item. The language used, the context of the problem and ease of understanding can affect the reliability and validity of the test as a whole. Some common mistakes that contribute to a reduction in validity and reliability include the following:

- Ambiguous questions, ie questions that have multiple interpretations;
- Bias items, such as items that are favour certain social backgrounds;
- The use of jargon that is not familiar to the target group.

Avoiding these mistakes should enhance the validity and reliability of the test scores.

Test Documentation

Once test items are constructed, they need to be assembled and documented for record and reproduction purposes. Apart from that, test documentation is also extremely important for evaluation and refinement purposes.

Test Evaluation

Test evaluation can be formative or summative. Formative evaluation can be carried out by administering a newly drafted test to a small group of students that is similar to the target group. The items are then analysed and the reliability of the scores are estimated. The results of this evaluation can be utilised to refine any test items found to be inadequate. A summative evaluation is performed in a similar manner but is based on the actual target group of students. In this case, test refinements can only be of benefit to the next batch of students.

Item Analysis

During the item analysis stage, a teacher can estimate the item quality indicators, specifically the item total correlation (ITC), which indicates the consistency of items, the difficulty index (Diff P) and items discrimination index (Disc D). These quality indicators can alert teachers to poor items. For example, an item that has a very high Diff P may be too easy. A Diff P of 0.5 is suitable for a norm-referenced test. An item that has low Disc D

Table 2: A table of specifications for a one-hour test on structural element design.

	COGNITIVE EMPHASIS			
CONTENT	Knowledge & Comprehension	Application & Analysis	Synthesis & Evaluation	Total (Content)
Beam design	10%	20%	20%	50%
-Load assessment	-	MC (4 @ 1 mark each)		
-Structural behaviour	SQ (2 @ 5 mark each)	SQ (2 @ 4 mark each)		
-Design		SQ(2 @ 4 marks each)	LQ (2 @ 10 marks each)	
Column design	10%	10%	0%	20%
-Axis of rotation	MC(2 @ 1 mark each)			
-Projection	MC (2 @ 1 mark each)			
-Plane of bending	MC (2 @ 1 mark each)			
-Effective height	MC (2 @ 2 marks each)	MC(2 @ 2 marks each)		
-Structural behaviour	-	SQ (1 @ 6 marks each)		
Slab design	0%	20%	10%	30%
-Load assessment on Slab	-	SQ (2 @ 4 marks each)		
-Structural Behaviour		SQ (2 @ 6 marks each)		
-Design	-		LQ (1 @ 10 marks each)	
Total (Cognitive Emphasis)	20%	50%	30%	100%

Table 3: Item analysis.

Quest	Correct	A	B	C	D	E	Omit	Total	Disc D	Diff. P
1-B	14	0	14	1	0	0	0	15	0.33	0.77
	9	1	9	1	2	2	0	15		

may not be discriminating between low and high achievers. A Disc D of 0.4 is considered adequate for classroom tests. By undertaking item analyses, teachers can identify some of the weaknesses of the items and thus improve upon them. The results of the try out test can then be used to refine the test.

An example of item analysis results using *Excel* and taken from Alias is given in Table 3 [5]. The example in shown in Table 3 is based on data taken from 30 students, and the formula for Disc D is given as Disc D = (U – L)/n, where U is the 50% of upper scores and L is the 50% lower score. The formula for Diff P is given as Diff P = (U+L)/Total N. Both formulae are from Black [3].

Estimating Reliability

In addition to the items analysis, a teacher should gain some estimate of the reliability of his/her test's scores as part of the evaluation process. The reliability for norm-referenced classroom tests may be estimated using various methods, with the Cronbach Alpha method being the most common method used. The Cronbach Alpha method provides estimates of internal consistency based on all possible split halves, while the split-half method provides an estimate of internal consistency based on two equivalent halves. The Cronbach Alpha coefficient, α, may be estimated using Equation 1,

$$\alpha = \frac{N}{N-1}\left[1-\frac{\sum_{i=1}^{N} S_i^2}{S_x^2}\right] \qquad (1)$$

where, *N is the n*umber of items (or identifiable parts of essay questions), S_i^2 is the variance of individual questions (or parts) and S_x^2 is the variance of whole test. An alpha coefficient of around 0.7 can be considered adequate for classroom tests. A lesser value may be obtained if heterogeneous items are used. The author managed to obtain a good reliability estimate of 0.74 based on the table of specifications shown in Table 2 [6].

Estimating Cronbach Alpha Coefficient Using Spreadsheets

The use of spreadsheets in estimating test reliability may greatly reduce the workload associated with repetitive hand calculations for teachers. Table 4 is the suggested spreadsheet format for reliability calculation adapted from Black [3]. The shaded cells constitute scores that the teacher has to key in while the rest are computed using formulae in *Excel*. Having once set up the spreadsheet, it is readily available for future usage.

Similarly, inter-scorer reliability can also be easily estimated using the template in Table 4. In this case, *Items* are replaced by *Examiners*, and the scores for each item in Table 3 (shaded cells) are replaced by total scores for each student.

Administering the Test

Tests that are well designed but not administered in an appropriate manner may still fail to produce reliable results, ie the results produced are not representative of students' actual capabilities. An example of instances where a test fails to be administered in an appropriate manner is when poor invigilation allows cheating among students. In this case, students' performance may be higher than actual scores and the results are not valid because there is inconsistency between actual and obtained scores. Therefore, even at this stage, care should be taken in order to avoid raising doubts over validity and reliability.

Table 4: The Alpha calculation for a norm-referenced test.

			Students					
Item	S1	S2	S3	S4	S5	S_1	S_1^2	means
1	63	37	73	76	59	13.8	190.2	61.6
2	53	33	68	78	60	15.2	230.6	58.4
3	64	46	75	88	68	13.8	189.8	68.2
4	53	45	65	86	61	13.8	191.2	62.0
5	52	44	67	87	69	14.9	221.4	63.8
Totals	285	205	348	415	317	Sum=	1023.2	
Mean =	314							
s_x =	69.409		N =	5				
s_x^2 =	4817.6		Alpha=	0.98				

CONCLUSION

Classroom tests are routinely designed and administered by teachers. To be of real value they must be valid and reliable. Test validity and reliability may be achieved from the steps taken throughout the design and administration stages. Two of the most effective methods that could be employed to enhance reliability and validity are constructing a table of specifications and carrying out a pilot study on the newly designed test. For increased efficiency, teachers may decide to work in teams to design and develop classroom tests. Lastly, although following the recommended measures previously discussed does not provide a guarantee for a perfect and valid test, it can certainly help teachers from getting it totally wrong.

REFERENCES

1. Gagné, R.M., *The Conditions of Learning and Theory of Instruction* (4th edn). New York: Holt, Rinehart and Winston (1985).
2. Alias, M and Gray, D.E, The learning hierarchy technique: an instructional analysis tool in engineering education. *Australasian J. of Engng. Educ.* (2005), http://www.aaee.com.au/journal/2005/alias_gray05.pdf
3. Black, T.R., *Doing Quantitative Research in the Social Science: An Integrated Approach to Research Design, Measurement and Statistics*. London: Sage (1999).
4. Bloom, B.S., *Taxonomy of Educational Objectives: The Classification of Educational Goals, Cognitive Domain*. London: Longmans (1956).
5. Alias, M., Spatial Visualisation Ability and Problem Solving in Civil Engineering. Unpublished Doctoral Thesis, University of Surrey (2000).
6. Alias, M., Black, T.R and Gray, D.E., The relationship between spatial visualisation ability and problem solving in structural design. *World Trans. on Engng. and Technology Educ.*, 2, **2**, 273-276 (2003).