African Educational Research Journal
Vol. 2(2), pp. 61-71, May 2014
ISSN: 2354-2160
Full Length Research Paper

Validity and reliability of teacher-made tests: Case study of year 11 physics in Nyahururu District of Kenya

Accepted 10 March, 2014

ABSTRACT

This study was carried out to establish the factors influencing the validity and reliability of teacher made tests in Kenya. It was conducted in Nyahururu District of Laikipia County in Kenya. The study involved 42 teachers and 15 key informants selected from teachers holding various positions of academic responsibilities in their schools in Nyahururu District. A mixed descriptive survey research design was applied. Data was collected through questionnaires and interviews with key informants. Quantitative data analysis was applied to survey data collected via questionnaires. The frequency distribution was described while data from interviews were qualitatively analyzed. The findings of the study revealed that the experience of teachers, training on test construction and analysis, level of education, use of Bloom's taxonomy, moderation of tests and length of tests have an effect on validity and reliability of the tests. Also these factors have a varying influence on the validity and reliability of teacher-made tests. Experienced teachers who had prior training in testing and therefore applied a number of these factors in their test construction tended to design tests with higher validity and reliability than their counterparts without such training. It was concluded that teacher-made tests are generally valid and reliable. The study recommended that teacher training on test construction and analysis needs to be enhanced in order to raise tests validity and reliability.

Keywords: Teacher-made tests, test design, validity, reliability.

*Corresponding author. E-mail: lukke7@gmail.com.

INTRODUCTION

In recent years, the quality of education has focused a great deal of attention on accountability. One of the ways by which such accountability is measured is by the extent to which students' performance in teacher-made tests can predict their potential performance in the standardized tests such as national examinations (Notar et al., 2004). Ideally, it is expected that there would be a strong positive correlation between a student's grade point average and the student's score on standardized tests (ibid). The grade point average is the mean grade score of all the formative assessment tests taken by the learner.

Formative assessment as implemented under the paradigm of Assessment for Learning (AfL) is considered a key aspect of teaching/learning process (Clark, 2008). This is driven by the assumption that if students are regularly informed about their progress in learning, then they are bound to learn better as a result of this feedback. In order to achieve this goal, regular formative assessments can be useful (Clark, 2008). In the course of the learning process, such formative assessments allow students to see their own progress and teachers to identify aspects of the content where more effective instruction is required (Tomlinson, 2008).

In contexts where students sit summative assessments such as national examinations for entry into tertiary learning institutions, the more effective the assessment becomes, the better the performance on the summative assessment is likely to be. This means that the effectiveness of the formative assessment can largely predict the outcome of the summative assessment (Hills, 1991).

One of the most cost effective ways of operationalising AfL, is the provision of teacher-made tests to learners as part of the learning process (Clark, 2008). Teacher-made tests are usually criterion referenced tests that are designed to assess student mastery of a specific body of knowledge (Wiggins, 1989). Unfortunately, studies and reviews on the impact of formative assessment on students' achievement in summative assessment have not been very positive.

A recent review of studies on this topic shows that a student's grade point average is usually not consistent with the same student's scores on standardized tests (Notar et al., 2004). Similar findings have been reported by Kingston and Nash (2011) in a meta-analysis whose conclusion was that formative assessment seems not to have a robust impact on students' achievement. This meta-analysis reported that the median effect size was only 0.20 instead of the 0.40 that often reported.

It has been argued that the problem of using such formative assessment for evaluation is that the teacher-made tests themselves are often severely flawed (Burton and Calfee, 1989). According to Wiggins (1989), "most criterion-referenced tests are inadequate because the problems are contrived and the cues artificial" (Wiggins 1989:708). It has been suggested that if teacher-made tests are going to adequately prepare the learners for the summative assessment at the end of the various key stages of learning, then teacher-made tests and end of key stage examinations must be comparable on the key attributes of test quality namely, validity and reliability (Parr and Bauer, 2006).

Validity and reliability of teacher-made tests

Validity

In the simplest terms, a test can be judged valid if it measures what it is intended to measure (Hathcoat, 2013). However, there is simmering controversy as per what validity in testing is with two schools of thought vie for dominance. On the one hand is the position that views validity as an attribute of score-based inferences and entailed uses of test scores while on the other, there is the instrument-based approach that holds that tests are either inherently valid or invalid.

This difference in meaning has influence on the reasons for validating score. That is, the question of the kind of evidence one ought to be looking for in the process of validation of a test arises out of these semantic differences. For example, to what extent does the observed difference in scores reflect the real underlying attribute (ibid). The instrument-based approach tends to be easily accepted in psychological testing because it means that there is a real attribute that is being measured. The conception of validity as an entity attributed to a test as a result of the manner in which the scores are interpreted tends to be given little value in a criterion referenced tests such as teacher-made tests where a criteria for scoring each item is preset. In the case of school tests, the criterion for what is the correct answer to an item is determined by available scientific knowledge about a phenomenon. There is only one interpretation given to the scores in a test, that is, a result of Test A is influenced by Attribute B (Borsboom, 2005).

The interpretation-based type of conception of validity tends to make sense in a norm-referenced test because the scores are interpreted as per the observed norm. That is, a test will just be a test and only the interpretations made about the test within a given norm is what is either valid or invalid (Hathcoat, 2013).

This means that both instrument-based and interpretation-based approaches to validity are applicable. However, the interpretation-based approach tends to have much broader application in that it can be used in virtually all contexts of testing, but the instrument-based testing is applicable in context where specific attributes are being measured (Hathcoat, 2013).

In the context of assessment at school, only certain attributes are targeted because learning objectives are usually specific and not general (Mager, 1997). Thus, the instrument-based approach to validity is what is best applicable to testing at school (Hathcoat, 2013). It is the accuracy of truthfulness of measurement vis-à-vis a given attribute as described in the learning objective based on the learner's performance (Hunter and Schmidt, 1999). For example, how one knows that a Mathematics test measures student's mathematical ability to solve mathematical problems not their reading skills. Whatever other factors that may have influenced the outcome of the test such as sickness during the test, not having time to do homework etc will not be accounted in the final judgement of validity of a test (Hathcoat, 2013).

Types of validity

There are different types of validity, that is, Face, Content, Criterion-related, and Construct validity. Face validity is where from a mere look at the test it is possible to deduce that the test is valid. This type of validity is not scientific though. For example, if a given test is supposed to measure mathematical skills, then by a mere fact that the items involve calculation in solving mathematical problems, then from the face of it such a test will be deemed valid.

Construct validity seeks to ensure that the test is actually measuring the intended attribute and not other extraneous attributes. For example, if a mathematics test is designed using difficult vocabulary beyond the level of the learner, that such a test will described as having low construct validity because it measuring other constructs besides the intended mathematical ones.

Criterion-related validity is of two types. Concurrent validity is where the results of one test are compared with those of another test across the same attribute. For example, the newer State of Anxiety Scale can administered at the same time as the older and much more established Taylor Manifest Anxiety Scale so that if the results of the former are comparable to the later, the former test will have passed criterion validity test. The other type of criterion-related validity is predictive validity. Here the performance of one test is used to predict the potential performance in another test. For example, the performance in an English test being used to predict how one will perform in mathematics.

Content validity or sampling validity ensures that a test covers broad areas of the syllabus. Items are sampled from across the syllabus and not just a specific topic. This facilitated by way of moderating a test using a panel to ensure that the designer does not just construct items testing the topics he/she likes only.

Formative validity seeks to establish the extent to which a test is able to provide information that can help improve the manner in which a program functions. For example, in Assessment for learning, the aim is to collect information that will improve the manner in which teaching is done for the benefit of the learner (Clark, 2008).

Criteria for evaluating validity of a test

Whatever the type of validity a tester is intending, Linn et al. (1991) proposed eight criteria for evaluating validity in performance-based assessment that cross-cut the above types of validity. These are the: (i) consequences, that is, on the effects of the assessment on the learner. The test constructor will be asking questions regarding intended purpose of test as and to what extent the learner is prepared to live by this purpose; (ii) content quality focuses on the consistency with current content conceptualization; (iii) transfer and generalizability focuses on the assessment's representatives of a larger domain; (iv) cognitive complexity focuses on whether the cognitive level of knowledge assessed is commensurate with the learner’s experiences; (v) meaningfulness addresses the aspect relevance of the assessment in the minds of students; (vi) Fairness deals with aspect of extent to which the test items are taking into account potential individual differences among learners; (vii) cost and efficiency focuses on the practicality or feasibility of an assessment in terms of the cost of producing and administrating the test and time required to complete the tasks.

Reliability

Reliability refers to the consistency of the scores obtained. That is, how consistent the scores are for each individual from one administration of an instrument to another and from one item to another. Reliability is a measure of how stable, dependable, trustworthy and consistent a test is in measuring the same thing each time (Worthen et al., 1993).

Factors that can simultaneously affect validity and reliability of a test

There are three variables that can affect the validity and reliability of teacher-made tests: the test taker, the environment and the test. It has been noted that the characteristic of the test-taker can affect the validity and reliability of the tests. Cassel (2003) has developed a testing method to determine the consistency and reliability of the test taker, a statistical measurement called a confluence score. This score looks at paired items in a test to show that the test taker is consistent in answering questions. Using confluence scores, the teacher would have to design the test so that a percentage of the questions would be asked seeking the same information in an opposite form. The student responses to these questions should be consistent. A student who gets one of these questions right and the other wrong is not a reliable test taker and should not be used to assess the validity of the test itself (Cassel, 2003).

The testing environment is another variable associated with the validity of teacher-made tests. If the testing environment is distracting or noisy or the test-taker is unhealthy, he or she will have a difficult time remaining consistent throughout the testing process (Griswold, 1990).

Even though actions ought to be taken to ensure that the testing environment is comfortable, adequately lit with limited interruptions (Griswold, 1990), these factor and the former one are largely aspects of test administrative procedures that are external to the test itself. This is because even in contexts where the characteristics of the test taker and the environment are well taken care of, it emerges that individual difference in performance will still be recorded.

This means that the third intrinsic variable affecting reliability and validity of teacher-made tests no matter the characteristics of the test-taker and the environment is the quality of tests themselves. The length of tests, use of Bloom’s taxonomy in test item construction and prior training of teachers on test construction to enable the teachers to design items that address various cognitive levels of thinking as per the Bloom’s taxonomy across the curriculum will all affect the validity and reliability of a given test.

The length or number of items is a crucial factor of test reliability. Carefully written tests with an adequate number of items usually produce high reliability (Justin and John, 1996) since they usually provide a representative sample of the behavior being measured and the scores are apt to be less distorted by chance factors, for example, familiarity with a given item or misunderstanding of what is expected from an item (Linn and Gronlund, 1995).

Long tests do three things to help maintain validity. Firstly, they increase the amount of content that the student must address, ensuring a more accurate picture of student knowledge. Secondly, long tests counteract the effects of faulty items by providing a greater number of better items. Third, long tests reduce the impact of student guessing (Griswold 1990).

In addition to the length of the test, there are several things to consider while trying to ensure the content is valid and reliable. First, test questions cannot be ambiguous. Poorly written questions will prompt students to guess, thus diminishing the reliability of the test. Second, test items need to be at a reasonable difficulty level (Griswold, 1990).

A report by Newell (2002) asserts that teacher-made tests usually measure only a limited part of a subject area, they do not cover a broad range of abilities and they rely too heavily on memorized facts and procedures. To guard any fortuitous imbalances and disproportionate item distribution, test constructors draws up a table of specifications of the target cognitive objectives as per Bloom’s taxonomy before any items are prepared. Such specifications as spelt out in Bloom (1956) should begin with an outline of both the instructional objectives of the course, the subject matter to be covered, and the cognitive skills measured (Gronlund, 1990). The time and effort expended to develop a table of specification can ensure that the test is valid and reliable (Notar et al., 2004).

Training in test construction is also an important factor. While some teachers report that they are confident in their ability to produce valid and reliable tests (Oescher and Kirby, 1990; Wise et al., 1991), others report a level of discomfort with the quality of their own tests (Stiggins and Bridgeford, 1985). Other teachers believe that their training in testing was inadequate (Wise et al., 1991). Indeed, most state certification systems and most of teacher education programs have no assessment course requirement or even an explicit requirement that teachers have received training in assessment. Instead, testing is taught as part of the foundational course of educational psychology (Boothroyd et al., 1992; Stiggins, 1991; Trice, 2000; Wise et al., 1991).

This formal training in assessment that teachers receive quite often focuses on large-scale test administration and standardized test score interpretation rather than on the test construction strategies or item-writing rules that teachers need for their own teacher-made tests (Stiggins, 1991; Stiggins and Bridgeford, 1985). Worse still, teachers have historically received little or no training or support after certification (Herman and Dorr-Bremme, 1984). One study by Mayo (1967) found that graduating seniors in 86 teacher-training institutions did not demonstrate a very high level of measurement competence.

Purpose

Therefore, this study sought to establish whether all other factors such as the characteristics of the test-taker and environment being equal, the tests that are designed by teachers are valid and reliable. Determine the factors affecting reliability and validity of teacher made tests. The study determined the extent to which each of the factors affects validity and reliability and their significance in determining validity and reliability.

METHODOLOGY

A descriptive survey research design was used to evaluate the validity and reliability of teacher made classroom tests in Nyahururu District.

Participants

A sample of 42 physics teachers was selected using stratified random sampling on the basis of their age-groups out of a population of 45 physics teachers in Laikipia County in Kenya. Of the total sample, 7.3% were teachers aged between 21 and 30 years, 51.2% were ranged between 31 and 40 years, 39% were aged between 41 and 50 years represented by 39.0%. Finally, those aged 51 to 60 were represented by 2.5%. Majority of the respondents (80.5%) were males with only 19.5% being females. This gender disparity was largely due to the fact that Laikipia County is a hardship area and so few female teachers are either posted here by the government or willing to come to in such an area.

Instruments

Questionnaires consisting closed ended questions were used to collect the primary data. They were administered by the researcher. The researcher also collected test papers for the term two of the year 2012 of the academic calendar of Kenyan school that runs between May to mid-August. The examination results of these tests were also collected for analysis of their validity and reliability. Structured interviews were used to get more in depth information from the teachers about their own opinion of their tests with regard to the aspects of validity and reliability.

Data collection procedure

The researcher personally visited the respective schools in which the teachers were sampled. After getting the school principal’s permission to conduct the study, he distributed the questionnaire to the sampled teachers to complete. Teachers were then requested to provide the researcher with the second term physics examination results for Year 11 class of 2012. Thereafter the researcher conducted face-to-face interviews with curriculum heads of the visited schools to establish their opinion about the validity and reliability of teacher made tests.

Data analysis

The data was coded and run through the statistical Program for Social Sciences (SPSS). The qualitative data was analyzed using

Table 1. Teachers experience and corresponding measures of validity and reliability.

Level of experience	V_1	V_2	V_3	V_4	V_5	V_6	V_7	Reliability (Kuder-Richardson method)
1	4	4	4	4	4	4	4	0.7
2	2	3	4	5	5	4	4	0.8
3	4	4	4	5	5	4	4	0.7
4	4	4	5	5	5	5	5	0.8
5	5	5	5	5	5	5	5	0.8

descriptive statistics. The researcher transcribed the questionnaire into the SPSS program for analysis.

The researcher with the help of two other teachers evaluated various examination questions and inter-rated them to determine their level in the criteria for validity. The inter-raters were teachers with experience of more than 10 years and have worked in more than one station. They were also experienced examiners in that they had been appointed by the Kenya National Examinations Council (KNEC) as examiners. The validity of the tests was evaluated across the various attributes that characterize validity such as the consequences for testing, the quality of content, generalizability of the results, cognitive complexity of the items, meaningfulness of the item content, to what extent the items are taking into account the backgrounds of the test-takers, that is, fairness and finally, the cost constructing the test and the efficacy of administering it. All these attributes were rated in the Likert scales using different adjectives as presented as follows:

V_1-consequencies: (1) very inconsequential (2) inconsequential (3) neutral (4) consequential (5) very consequential
V_2-content quality: (1) very low content quality (2) low content quality (3).moderate content quality (4) high content quality (5) very high content quality.
V_3-generalizability: (1) very specific (2) specific (3) neutral (4) generalizable (5) totally generalizable.
V_4-cognitive complexity: (1) very simple (2) simple (3) fair (4).complex (5) very complex
V_5- meaningfulness: (1) very meaningless (2) meaningless (3) neutral (4) meaningful (5) very meaningful.
V_6-fairness: (1) very unfair (2) unfair (3) neutral (4) fair (5) very fair.
V_7-cost and efficiency: (1) very expensive and inefficient (2) expensive and inefficient, (3) neutral (4) cheap and efficient (5) cheap and efficient.

Teachers' level of experience was categorized according to Dreyfus and Dreyfus (1980) of each factor was also classified as follows:

1. 1-5 years - Novice who had little situational perception and discretional judgment.
2. 6-10 years - Advanced beginner with all attributes and aspects treated separately.
3. 11-15 years - Competent whose plan guides performance as situation evolves.
4. 16-20 years - Proficient with situational factors guiding performance as situation evolves.
5. > 20 years-Expertise with intuitive recognition of appropriate decision or action.

After the above classifications for each of the factors, the results were analyzed and averaged to get the value of each exam.

The quantitative data was analyzed using quantitative methods and presented by use of tables, frequencies, percentages, statistical measures of relationship between the dependent and independent variables. The researcher also carried out an analysis of examination results in order to calculate the reliability coefficient. The results were used to draw conclusions and in making recommendations.

RESULTS

Factors affecting validity and reliability

This study was interested in determining the factors that affect reliability and validity of teacher made tests. It was established that there are various factors that affect the reliability and validity of teacher made tests.

Teachers' experience

From Table 1, it is shown that as the experience of teachers increase there is a corresponding increase in validity.

Though teachers experience affects and varies with the number of years one has been teaching, this does not seem to have the same effect on reliability. This implies that validity is affected so much by the experience of teachers rather than reliability.

Table 2 shows a summary of the number of years that teachers have been teaching. Teachers were asked how many years that have been teaching.

Most of the respondents have an experience of 16 to 20 years (31.7%), followed closely by those with11 to 15 years (24.5%), then those with1 to 5 years (19.5%), those with more than 20 years follow with 17.1% and finally those with 6 to 10 years at 7.3%. This shows that most teachers have an experience of between 11 and 20 years. This is good experience to be able to perform their responsibility of testing effectively.

The key informants were asked whether teachers experience affects the reliability and validity of the tests they construct. The following is an excerpt from of discussion with few interviewees:

Interviewer: Does the number of years one has taught affect quality of their test in terms of validity and reliability?
Key informant 1: Definitely it does.
Key informant 2: Yes it does, because initially it is kind if they are still in training.
Key informant 3: Yes it does and keeps improving all

Table 2. Distribution of respondents by experience.

Number of years	Frequency	Percentage
1-5	8	19.5
6-10	3	7.3
11-15	10	24.5
16-20	13	31.7
>20	7	17.1
Total	41	100.0

Table 3. Education level versus levels of reliability and validity.

Levels	V_1	V_2	V_3	V_4	V_5	V_6	V_7	Reliability (Kuder-Richardson method)
1	-	-	-	-	-	-	-	-
2	4	4	4	4	4	4	4	0.68
3	5	5	5	5	5	5	5	0.74
4	-	-	-	-	-	-	-	-
5	-	-	-	-	-	-	-	-

Table 4. Distribution of respondents by their education level.

Level of education	Frequency	Percentage
Certificate	0	0.0
Diploma	11	26.8
Bachelors	30	73.2
Masters	0	0.0
Doctorate	0	0.0
Total	41	100.0

along.
Key informant 4: It is true it does

Most of them were of the opinion that as the number of years increases, the validity and reliability will also improve though also it has a limit kind of where one attains optimum point. The $\chi^2 = 6.683$, with p=0.01 and 4 d.f. Thus the significance of experience is quite high.

Education level

Table 3 shows that the level of education affects reliability and validity. The respondents were asked to state their level of education.

As the level of education rises also validity and reliability also get better. This shows that the level of education affects reliability and validity of teacher made tests.

The distribution of respondents in terms of their education level is shown in Table 4. The respondents were asked to state their level of education, that is, Certificate, Diploma, Bachelors, Masters or Doctorate.

Most of the respondents have Bachelors degree (73.2%) while others have Diploma (26.8%). Most of the teachers are well qualified to set tests that meet the criteria for validity and also have a high reliability. The key informants were asked whether they thought that education level affected quality of teacher made tests. The following is an excerpt of their responses:

Interviewer: Education level affects validity and reliability of teacher made Tests.
Comment.
Key informant 1: It does affect and the more qualified one is the better.
Key informant 2: To a great extent it does.
Key informant 3: Yes it does affect but not so much.
Key informant 4: It depends on the level you are comparing. If between certificate and degree there is a lot of difference but not so big between diploma and degree holders.
Key informant 5: Yes it does.

Most of the key informants felt that the education level does affect the level of reliability and validity. It was felt

Table 5. Training on test construction and corresponding values of reliability and validity.

Level of training	V_1	V_2	V_3	V_4	V_5	V_6	V_7	Reliability (Kuder-Richardson method)
1	3	3	3	3	3	3	3	0.64
2	3	3	3	3	3	3	3	0.68
3	3	3	3	3	3	3	3	0.76
4	5	5	5	5	5	5	5	0.8
5	4	4	5	4	4	5	4	0.79

Table 6. Summary of the results on training on test construction and analysis.

	Frequency	Percentage
Strongly disagree	21	51.2
Disagree	1	2.4
Do not know	6	14.6
Agree	12	29.3
Strongly agree	1	2.4
Total	41	100.0

Table 7. Distribution of respondents on need for further training on test construction and analysis.

	Frequency	percentage
Strongly disagree	0	0
Disagree	3	7.3
Neutral	0	0
Agree	13	31.7
Strongly agree	25	61.0
Total	41	100.0

that one could not take others to where they have never been before. The better one is trained therefore, the more efficient they become in constructing valid and reliable tests. The $\chi^2 = 8.805$ at $p = 0.01$ significance and df = 4 of freedom, which shows that the level of training is significant in determining validity and reliability of teacher made tests.

Training on test construction and analysis

This study was interested in identifying whether training on test construction and analysis has effect on validity and reliability.

From Table 5 as the level of training one have on test construction increases also the reliability increases. The trend is observed when we consider reliability though there is no much variation. When asked whether the training one has had on test construction and analysis is adequate most of the respondents strongly disagreed (51.2%) while only 2.4% strongly agreed that their training is adequate. Table 6 shows a summary of the results.

When the respondents were asked whether they need further test on test construction and analysis most of them strongly agreed (61%), while 31.7% agreed. Only 7.3% disagreed and none was neutral or strongly disagreed. Table 7 shows a summary of the results. The $\chi^2 = 34.976$ at 0.01 significance and df = 4. There is high significance of training on test construction and analysis.

Most of the respondents are not satisfied in their current level of training in test construction and analysis to further enhance their skills.

The key informants were asked whether they think teachers have enough training on test construction and whether more training is needed. The following is an excerpt from a few key informants:

Interviewer: Do you think teachers are well trained on test construction or more is needed.
Key informant 1_1: There is a lot that needs to be done on training of teachers on test construction and analysis.
Key informant 2: Even the institutions from which teachers come from affect the level of reliability and

Table 8. Usage of Bloom's taxonomy and corresponding values of reliability and validity.

Level of usage of Bloom's taxonomy	V_1	V_2	V_3	V_4	V_5	V_6	V_7	Reliability (Kuder-Richardson method)
1	3	3	3	3	3	3	3	0.6
2	3	3	3	3	3	3	3	0.60
3	4	4	4	4	4	4	5	0.7
4	4	4	4	4	4	4	4	0.79
5	4	4	4	4	4	4	4	0.82

Table 9. Distribution of respondents in terms of use of Bloom's taxonomy.

	Frequency	Percentage
Strongly disagree	1	2.4
Disagree	4	9.9
Neutral	24	58.5
Agree	11	26.8
Strongly agree	1	2.4
Total	41	100.0

reliability and a lot needs to be done to harmonize these disparities.

Key informant 3: The training leaves a lot to be desired. Teachers need a lot of in-service training on test construction and analysis.

Key informant 4: All the teachers need to keep sharpening their skills in test construction and analysis.

Most of the key informants agree strongly that teachers need more training on test construction and analysis. When the question was posed to them the key informants also felt that they themselves need to be trained on test construction and analysis. Some went further to comment that there is need for more training when one undergoes the teaching course in teacher training institutions. The $\chi^2 = 17.756$ at 0.01 significance level and df = 4 which indicates the importance placed on further training on test construction.

Use of Bloom's taxonomy specification table

The use of Bloom's taxonomy in constructing tests is more of a rule rather than exception. Table 8 shows that use of bloom's taxonomy improves the reliability and validity of the teacher made tests.

Those exams set in accordance to the levels of Bloom's taxonomy have a high reliability and also validity. The only surprising thing is that most teachers never really seem to be sure whether they use Bloom's taxonomy. The respondents were asked whether they use Bloom's taxonomy in the process of constructing their exams. Table 9 is a summary of the results obtained.

The key informants could not be further from the truth since when asked whether teachers use Bloom's taxonomy most were not sure about. The excerpt below is a summary of some responses from key informants.

Interviewer: Do teachers use Bloom's taxonomy in preparing tests.

Key informant 1: Not really. Most teachers do not seem to understand it.

Key informant 2: No. Teachers are not able to identify objectives in there tests using Bloom's taxonomy.

Key informant 3: Not really. Teachers do not seem to appreciate the use of Bloom's taxonomy or recognize its importance when preparing tests.

Indeed most said that they did not really think that teachers used Bloom's taxonomy. It seems either there are no mechanisms of determining whether teachers use this taxonomy or there is a lack of expertise to do so. Also most would say it takes a lot of time to construct an exam using Bloom's taxonomy. The $\chi^2 = 46.195$ at p = 0.01 and df = 4. There is great importance in using Bloom's taxonomy to raise the value of validity and reliability.

Moderation of the tests

Table 10 shows that the moderated exams had a higher validity and also a high reliability. This shows that the input of the members of department is very important in

Table 10. Moderation of tests and corresponding values of validity and reliability.

Level of moderation	V_1	V_2	V_3	V_4	V_5	V_6	V_7	Reliability (Kuder-Richardson method)
1	4	4	4	4	3	4	4	0.8
2	3	5	4	3	5	5	2	0.85
3	5	5	5	5	5	5	3	0.85
4	5	5	5	5	5	5	5	0.89
5	5	5	5	5	5	5	5	0.8

Table 11. Distribution of respondents in terms of exam moderation.

	Frequency	Percentage
Strongly disagree	10	24.4
Disagree	6	14.6
Neutral	9	22.0
Agree	12	29.3
Strongly agree	4	9.9
Total	41	100.0

Table 12. Length of tests and level of validity and reliability.

Number of items	V_1	V_2	V_3	V_4	V_5	V_6	V_7	Reliability (Kuder-Richardson method)
1	3	3	3	3	3	3	3	0.6
2	3	4	3	5	4	3	3	0.68
3	3	5	4	5	4	3	4	0.8
4	3	4	4	4	4	4	4	0.8
5	5	5	5	5	5	5	5	0.8

improving the validity and reliability of the tests constructed by teachers.

The respondents were asked whether they moderate their tests. Table 11 is a summary of the results.

It seems that not many respondents subject their exams to moderation and this goes to affect the level of reliability and validity of teacher made tests.

Most of the key informants accepted that their respective institutions do not moderate exams but it is very important. The following is an excerpt of the discussion with a few key informants:

Interviewer: Is moderation important and do you do it in your institution.
Key informant 1: It is very important though we do not carry it out.
Key informant 2: Very important but we do not have time to do it.

The key informants reported that this never happens and most teachers assume that if they take questions from past papers and mix them with their own it serves to moderate the exams. Also teachers view that moderation is an unnecessary burden and yet they are trained to construct exams in universities and colleges. The χ^2 = 4.976 at 0.01 significance and df = 4. This shows that not a lot of interest was placed on moderation of tests. Indeed most teachers actually show less importance to moderation of test because they that it is their capability of being able to set good examinations that is actually being tested and not just the desire to improve the quality of the tests.

Length of the tests

As the length of a test increases the validity and reliability also do improve. This is shown in Table 12.

The respondents were asked how many items they normally use in their exams. Table 13 shows the results obtained for the number of items used in an examination by the respondents.

Most of the respondents construct tests with 16 to 20 items (65.9%), followed by 11 to 15 items (31.7%) and

Table 13. Results for the number of test items used by respondents.

	Frequency	Percentage
1-5	0	0.0
6-10	1	2.4
11-15	13	31.7
15-20	27	65.9
>20	0	0.0
Total	41	100.0

finally 6 to 10 items (2.4%). The key informants also felt that the number of items have an effect on the validity and reliability of a given test. When asked whether the number of items affect validity and reliability the key informants had a plain answer of yes it does. Most of them felt that the number items served to remove or to a great extent reduce the examiners biases and also encourage the learners. Though this is true, it is also better to consider the process of coming up with the questions so that it will not be just for the sake of having many items in the tests. $\chi^2 = 24.78$ at 0.01 and d.f. = 4. This is an indicator that most teachers place a lot of importance on the length of test as a factor that increases validity and reliability.

DISCUSSION

This study found that the quality of teacher made tests in terms of their validity and reliability is affected by a number or factors. The teachers with more experience prepared tests which were more valid and reliable. Findings of a study by Magno (2003) showed that highly experienced teachers prepared examinations with high validity and reliability. Similarly, teachers who are trained on test construction and analysis prepared tests that were more valid and reliable. According to Stiggins (1994), the level of education of the teacher usually affects the reliability and validity of teacher-made tests. This effect of training on the quality of tests also spread to other aspects of testing as noted by Marso and Pigge (1988) who argue that lack of planning for good tests is due to lack of training.

Another factor that influenced the quality of tests is the moderation of the tests prior to administration. Moderated tests have a higher level of reliability and validity than unmoderated ones. In this study, moderation of tests had a $\chi^2 = 4.976$ at 0.01 significance and d.f. = 4 and therefore proved to be an important factor affecting validity and reliability. In addition to the above, the application of Bloom's taxonomy was found to have a direct influence on validity and reliability. Teachers who used the table of specification to design test items generated tests with higher validity and reliability than those who did not. This was supported by the findings by Linn and Gronlund (1995) who recommended that it was important to plan a test using the table of specification such as Bloom's taxonomy in order to ensure proper sampling of items to meet conditions of validity and reliability. Recent research in this area supports the use of table of specification in test construction as a way of improving quality (Fives and DiDonato-Barnes, 2013). Finally, the length of the tests also affected the quality of the tests. According to Wells and Wollack (2003), longer tests produce higher reliabilities and validities. Indeed most teachers construct exams with many items in order to increase their reliability and validity and this was observed in this study. In this study, teachers who considered these factors well in preparing tests, their tests they have a high level of reliability and validity.

CONCLUSION AND RECOMMENDATIONS

A number of factors tend to influence the validity and reliability of teacher-made tests. In line with previous research, these factors range from lack of commitment to good practice in testing to lack of proper training in testing. While teachers who are trained and educated in tested tend to have a better understanding and practice when it comes to testing those without training and education tend to misunderstand attributed of good practice such as the purpose of moderation of examinations. The authors of this paper therefore recommended that teachers should regularly refreshed with in-service training in testing to ensure good practice with regard to the construction of teacher-made tests.

REFERENCES

Bloom, B. S. (Ed.). (1956). Taxonomy of educational objectives: Handbook I: Cognitive domain. New York: David McKay Company.

Boothroyd, R. A., McMorris, R. F., and Pruzek, R. M. (1992). What do teachers know about measurement and how did they find out? Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA. (ERIC Document Reproduction Service No. ED351309)

Borsboom, D. (2005). Measuring the mind: Conceptualissues in contemporary psychometrics. New York: Cambridge University Press.

Cassel, R. (2003). Confluence is a primary measure of test validity and it includes the creditability of test taker. College Student Journal, 37:348-353.

Clark, I. (2008) Assessment is for learning: Formative assessment and positive learning interactions. Florida Journal of Educational Administration and Policy, 2(1):1-16.

Dreyfus, S. E., and Dreyfus. H. L. (1980). A Five Stage-Model of the Mental Activities involved in Direct Skill Acquisition. Washington, DC: Storming Media.

Fives, H., and DiDonato-Barnes, N. (2013). Classroom test construction: The power of a table of specification. Practical Assessment, Research and Evaluation, 18(3):1-7

Griswold, P. A. (1990) Assessing relevance and reliability to improve the quality of teacher-made tests. NASSP Bulletin, 76:18-24.

Hathcoat, J. D. (2013). Validity semantics in educational and psychological assessment. Practical Assessment, Research and Evaluation, 18(9):1-14.

Herman, J. L., and Dorr-Bremme, D. W. (1984). Teachers and testing:

Implications from a national study. Draft. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA. (ERIC Document Reproduction Service No. ED244987)

Hills, J. R. (1991). Apathy concerning grading and testing. Phi Delta Kappan, 72(7):540-545.

Hunter, J. E., and Schmidt, F. L. (1990). Methods of meta-analysis: Correcting error and bias in research findings. Newbury Park, CA: Sage Publications.

Kingston, N., and Nash, B. (2011). Formative assessment: A meta-analysis and a call for research. Educational Measurement: Issues and Practice, 30(4):28-37.

Linn, R. E., Baker, E. L., and Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. Educational Researcher, 20(8):15-21.

Linn, R. L., and Gronlund, N. E. (1995). Measurement and assessment in teaching. New York: Prentice Hall.

Mager, R. F. (1997). Measuring instructional results: How to find out if your instructional objectives have been achieved. (3rd ed.). Atlanta, GA:CEP Press.

Magno, C. (2003). The profile of teacher-made test construction of the professors of University of Perpetual Help Laguna. UPHL Institutional Journal, 1(1):48-55.

Marso, R. N., and Pigge, F. L. (1988). An analysis of teacher-made tests: Testing practices, cognitive demands, and item construction errors. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA. (ERIC Document Reproduction Service No. ED298174).

Mayo, S. T. (1967). Pre-service preparation of teachers in educational measurement. U.S. Department of Health, Education and Welfare. Washington, DC: Office of Education/Bureau of Research.

Newell, R. J. (2002). A different look at accountability: The EdVisions approach. Phi Delta Kappan, 84:208-211.

Notar, C. E., Zuelke, D. C., Wilson, J. D., and Yunker, B. D. (2004). The table of specifications: Insuring accountability in teacher made tests. Journal of Instructional Psychology, 31:115-129.

Notar, C. E., Zuelke, D. C., Wilson, J. D., Yunker, B. D. (2004). The table of specifications: Insuring accountability in teacher made tests. Journal of instructional Psychology, June, 2004.

Oescher, J., and Kirby, P. C. (1990). Assessing teacher-made tests in secondary math and science classrooms. Paper presented at the annual meeting of the National Council on Measurement Education, Boston. (ERIC Document Reproduction Service No. ED 322 169)

Parr, A. M., and Bauer, W. (2006). Teacher made test reliability: a comparison of test scores and student study habits from Friday to Monday in a high school biology class in Monroe County Ohio. Masters Thesis. Graduate School of Marietta College. Access on 24th January, 2013 from: http://etd.ohiolink.edu/sendpdf.cgi/Parr%20Anita.pdf?marietta1142864088

Stiggins, R. J. (1991). Assessment literacy. Phi Delta Kappan, *72*:534–539*.*

Stiggins, R. J. (1994). Performance assessment. ERIC/CASS Digest Series on Assessment in Counseling Therapy.

Stiggins, R. J., and Bridgeford, N. J. (1985). The ecology of classroom assessment. *Journal* of Educational Measurement, *22*:271–286.

Tomlinson, P.D. (2008) Psychological theory and pedagogical effectiveness: The learning promotion potential framework. British Journal of Educational Psychology, 78(4), 507-526

Trice, A. D. (2000). A handbook of classroom assessment. New York: Addison Wesley Longman.

Wells, C. S., and Wollack, J. A. (2003). An Instructor's Guide to Understanding Test Reliability. Testing & Evaluation Services. University of Wisconsin. November 2003. Available at:http://testing.wisc.edu/Reliability.pdf. Retrieved March 28, 2013.

Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. Phi Delta Kappan, 70:703-710.

Wise, S. L., Lukin, L. E., and Roos, L. L. (1991). Teacher beliefs about training in testing and measurement. Journal of Teacher Education, 42:37–42.

Worthen, B. R., Borg, W. R., and White, K. R. (1993). Measurement and evaluation in the schools. White Plains, NY: Longman.