Making Effective Teacher-Made Tests of English:

A Major KPI for Thai Teachers of English

Dr. Supanrigar Watthanaboon[1] and Assist. Prof. Dr. Preemon Nakarin[2]

Faculty of Education, Suratthani Rajabhat University[1] and Walailuk University[2]

E-mail: supanrigar@yahoo.com[1] and npreemon@wu.ac.th[2]

(Received 4 May 2007; Accepted 29 July 2007)

Introduction

Central to school's evaluation process are teacher-made tests. Such instruments are designed to appraise the outcomes of local classroom instruction. Generally, commercial standardized tests are too general in scope and too inflexible to meet the special requirements of each subject or group of students, plus other diversified elements involved.

Experienced educators and teachers know that "good" tests do not simply happen. Nevertheless, test construction all too often occurs at the last possible moment and in haste. This is unfortunate, since testing is an integral aspect of the total teaching-learning program. In preparing a test, the teacher needs to have a clear conception of how the test, together with the test results are to be functioned and used; and requires prior specification of instructional objectives and decisions regarding the sequence and method of instruction.

Educational and psychological testing can help individuals at all levels of schooling make better decisions. Testing data may be employed in placing students, formative evaluation, diagnostic evaluation, making selection decisions, arriving at curricular decisions, personal decision making, and for summation purposes.

Consequently, it is very essential that the teachers, especially the Thai teachers who are involved and concerned in teaching ESP to Thai students, must have a good quality of competency, or KPI, in constructing effective teacher-made tests of the ESP courses taught and learned in schools of any level to competently evaluate the learning outcomes of the learners more effectively, by integrating their instruction and evaluation as one phase in an encompassing web of continual classroom planning.

The Quality of English Teacher-Made Tests

A measurement device used in teaching-learning process should possess several qualities. Among the most important of these are reliability, validity, and practicality (Gronlund and Linn, 1990). ***Reliability*** has to do with how accurately a measuring device, a test, and measures what the teacher sets out to measure and the precision of the resulting score. ***Validity*** concerns whether a

test measures what the teacher wants it to describe, represents all the components of what the teacher wants to describe, and describes nothing else but what we want it to describe. ***Practicality*** regards the ease with which a test can be administered and scored. A good many statistical procedures are available to educators and teachers for evaluating these properties of this kind of teacher-made tests or any test. Here, we shall consider what each quality involves, although a statistical treatment is usually reserved for more advanced courses in education and psychology.

Reliability

In deciding upon or designing a measuring device, we are concerned with how accurately it measures what we intend to measure and the precision of the resulting score. Thus, we need to know whether it will yield a similar result under similar conditions if we again measure the property in which we are interested.

Reliability refers to the degree to which an instrument yields a consistent measurement of the same thing. For instance, if we take our temperature and the thermometer registers readings varying from 95 degree to 103 degree F when we are known to have a normal temperature (which is around 98 degree), we will have little confidence in the instrument. Nor will we feel easy with the measurement of a house taken by an elastic ruler.

We confront the problem of reliability when we administer an achievement test to our students. Would the students realize the same scores if they took the test last week, yesterday, tomorrow, or next week? Would their scores be the same had we provided another test with a differing sample of what we believe to be equivalent items?

These matters deal with how generalizable test results are over different occasions or over different samples of the same type of behavior.

A good many factors affect the reliability of a measuring device (Ebel and Frisbie, 1990). The individuals taking the test may themselves change from one time to the next. Such changes include state of health, motivation, fatigue, emotional strain, attention, forgetting, guessing, and training. The simple fact of having previously taken a similar test also introduces change.

Further, the task itself may change, since the second test usually contains somewhat different items. And finally, the test administrators may not adhere to the time limit rigidly or the scorer may not grade the tests in the identical fashion (especially those with essay-typed items). Such of these factors introduce some element of error to all test scores (McKeachie, 2006).

Validity

Perhaps the most important question we have regarding a measuring device is whether it measures what we want it to describe, represents all the components of what we wish to describe, and describes nothing else but what we want it to describe. The matter is similar to the rule that the courtroom witness tell the truth, the whole truth, and nothing but the truth. The extent to which an instrument serves the purpose for which it is intended is termed ***Validity.***

If we are interested in the length of a desk, it is of little use to have a scale that determines its weight; instead, we require a ruler. Should we be interested in determining the language achievement of a group of students, we would need to prepare a test that adequately samples a variety of language skills. The score would represent a measure of each student's language proficiency. However, the score itself is not the student's proficiency but merely a record of a sample of the learner's behavior. Any appraisal regarding the students' proficiency is an inference from the number of problems the learners solve correctly. The validity of the score is not self-evident but must be established on the basis of adequate evidence (Genesee and Upshur, 2001). Three basic types of ***validity*** are to be identified: ***Content Validity, Criterion-Related Validity,*** and ***Construct Validity.***

Content Validity refers to ***the extent to which a test measures a representative sample of the subject matter content or the behavioral changes*** in which the ESP teachers are interested. Building the content validity of a test is equivalent to ascertaining how well it samples certain types of subject matter or behaviors. If the teachers are concerned with the vocabulary comprehension of a group of students, we would need to measure each student's performance on a sample of questions intended to represent an aspect of word learning achievement.

Criterion-Related Validity refers to ***the extent to which test performance is related to some other external measure.*** The teachers are, in effect, asking with what confidence they can generalize or predict from these test results how well a student will do on a ***different*** task. For example, a test may be used to estimate a student's ***present language skill.*** Thus, a dictation test may be interpreted as telling the teacher about the accuracy with which a student can perform the necessary dictation from the boss in a company's office. This type of validity of the test can then be assessed by how well the student actually take dictation assigned to him in the office setting.

Or a test may be employed to make a ***prediction about a student's future achievement.*** Colleges commonly use academic aptitude tests as part of their admission procedures. The tests are designed to forecast the probability of a student's college success. The validity of the test can then be experimentally determined by administering the test to a group of upper secondary school students, and then later assessing how well the test predicted these same students' grades at the end of their first year in college.

Construct Validity refers to ***the extent to which some hypothetical trait is reflected in the test performance.*** Various psychological and educational tests seek to measure general traits (constructs) like a person's verbal fluency, comprehensive skill, communicative ability, reasoning ability, spatial visibility and anxiety, and so forth. Tests of these qualities are considered valid insofar as they reveal the traits being expressed in the way that our existing body of knowledge says such traits should be expressed. For instance, from what we know regarding

assertiveness we would expect that a group of sales personnel should score especially high on a measure of assertiveness and a group of librarians should score low.

Practicality

In selecting or devising a test, practical considerations need to be taken into account. A primary consideration is ***the ease with which the test can be administered. The directions should be complete, simple, and clear.*** They should be in written form. The more complicated the directions and the greater the number of subtests, the more likely errors will occur that distort the results.

The scoring of tests has traditionally been a particularly tedious, cumbersome, and troublesome operation. However, the trend toward practical objective tests, the availability of separate answer sheets, and machine scoring have considerably eased many of the teacher's burdens.

The practicality of a test is also dependent upon the ease with which the results can be interpreted and applied to further instruction, solving, and correcting classroom learning problems, such as diagnosing student weaknesses, structuring remedial instruction, organizing class groupings, and things of this nature.

Functional English Teacher-Made Tests

Teachers are often as concerned with measuring the language ability of students to think about and use knowledge as they are with measuring the knowledge their students possess. In these instances, tests are needed that permit students some degree of freedom and diversity in their responses to the test questions. There are various types of teacher-made tests which have been formally in use in measuring the teaching-learning outcomes effectively in schools of all levels (Jacobs and Chase, 1992); namely, ***essay tests, short-answer tests, matching tests, True-False tests,*** and ***multiple-choice tests.***

Essay Tests

Some teachers claim that ***Essay Questions*** have a desirable effect on students' study habits (Ericksen, 1995). The questions of this type of test compel students to consider larger units of subject matter rather than preoccupying themselves with many isolated bits and pieces of knowledge. The test provide items/questions in which students supply, rather than select, the appropriate answer. Usually, the students compose a response in one or more sentences, Essay tests allow students to demonstrate their ability to recall, organize, synthesize, relate, analyze, and evaluate ideas, and all of these good learning skills must reflect in their answers to the essay questions. The ***major advantage*** is that the essay tests provide students with an opportunity to integrate and apply their thinking and problem solving skills creatively (Crooks, 2001). Rather than simply selecting a correct response, the student must supply an appropriate answer. As such, essay tests can provide an effective instrument for tapping higher level of reasoning.

Constructing essay test questions. In preparing essay questions, teachers commonly find it helpful to keep the following suggestions in mind.

1. ***Phrase the question with sufficient specificity so students know what they are asked to do.*** Avoid vague question with ambiguous wording. For example:

Poor: What is a cheque?

Better: Explain the definition of a "cheque" and its "function" in business.

2. The question should be written in a way that will elicit the desired response in terms of objectivity and evidence. This is especially important in asking students a question dealing with a controversial issue. Asking students "what is your opinion" or "what do you think" provides ***no basis*** for arriving at a generally acceptable answer. Instead, students ***should be asked to marshal evidence and arguments in support of one or another position.*** For example:

Poor: What is your opinion regarding the alienation business act?

Better: Considering pros and cons regarding the enactment of a new alienation business act, you are asked to outline evidence and arguments either in support of or in opposition to the enactment of the said new act.

3. When possible, ***phrase a question in a novel manner***. For example:

Poor: Explain the effect of a meander upon the banks of a river.

Better: You are planning to purchase land along a meandering river. Would it be better to purchase land on the inside or outside bank of a meander? Give the reasons for your good choice.

Objective-Item Tests

Objective-item tests are of two types. The ***supply type*** asks the student to provide a short answer or to complete a blank. The ***select type*** provides the student with alternative responses in the form of ***matching, true-false, or multiple-choice*** items. Proponents of objective-item tests contend that they assure good content sampling and easy and reliable scoring. Critics say that the tests foster rote learning, encourage guessing, and neglect the cultivation of integrating and organizing skills.

Short Answer Tests

Short answer item tests are of two types:

A. Simple Direct questions (e.g. ***Who was the first president of the United States?***) and

B. Completion Items (e.g. ***The name of the first president of the United States is ……..***)

These short answer items can be answered by a word, phrase, number or symbol. The short answer test is a cross between essay and objective tests. The student must supply the answer as with an essay question but in a highly abbreviated form as with an objective question.

Short-answer items have a number of advantages. First, they reduce the likelihood that a student will guess the correct answer. Second, they are relatively easy for a teacher to construct. Third, they are well adapted to mathematics, the sciences, and ***foreign languages*** where specific types of knowledge are tested. Fourth, they are

consistent with the logical question-and-answer format as straight forward to the point, ***no tricks.***

Matching Item Tests

The matching item test consists of ***two parallel columns. The column on the left contains the questions to be answered,*** termed ***premises; the column on the right, the answers,*** termed ***responses.*** The student is asked to associate each premise with a response to form a matching pair. For example:

Capital City	***Nation***
……. 1. Paris	*a. Denmark*
…….2. Copenhagen	*b. Spain*
…….3. Lisbon	*c. Portugal*
…….4. Madrid	*d. France*
……5.The Hague	*e. Netherlands*
	f. Hungary
	g. Germany

In some matching tests ***the number of premises and responses are the same,*** while in many others the premises and responses may be different as illustrated above.

The chief advantage of matching tests is that a good deal of factual information can be tested in minimal time, making the tests compact and efficient. They are especially well suited to ***who, what, when,*** and ***where types of subject matter.*** Moreover, students frequently find the tests fun to take because they have puzzle qualities to them.

The principal difficulty with matching tests is that teachers often find that ***the subject matter is insufficient in quantity or not well suited for matching items.*** The test of this type should ***be confined to homogeneous items containing one type of categorization of the subject matter,*** for instance, authors – novels; inventions – inventors; major events – dates; terms – definitions; foreign words – English words equivalents; rules – examples; and the like.

True – False Items Tests

The true-false item tests consist of a declarative statement that ***the students are to read and judge the given statements or items to be either correct or incorrect.*** Each question contains only two possible answers. Teachers find that the true or false items are easy to construct and score (SEAMEO, 2003), and that even students who are rather poor readers can cope with them.

However, the true-false items found on this type of test are too often focus upon unimportant pieces of information. The chief exceptions have to do with questions distinguishing between facts and opinion, and in identifying cause-and-effect relationships. Further, since there are only two alternatives, the students have a fifty-fifty opportunity of guessing the correct answer on chance alone.

Multiple-Choice Items Tests

The multiple-choice question is probably the most popular as well as the most widely applicable

and effective type of objective tests (Mehrens and Lehmann, 1991). It consists of two parts: (1) the ***stem, which states the problem or question,*** and (2) a list of ***three to five alternatives, one of which is the correct or best answer,*** and the others function as "distractors" or incorrect options that draw the less knowledgeable students away from the correct response. ***The stem may be stated as a direct question or as an incomplete statement***. For example:

Direct Question:

What is the capital city of Denmark?

a. Paris
b. Lisbon
c. Copenhagen
d. Rome

Incomplete Statement:

The capital city of Denmark is

a. Paris
b. Lisbon
c. Copenhagen
d. Rome

The chief advantage of the multiple-choice tests is its versatility. For instance, it is capable of being applied to a wide range of subject areas. In contrast, short-answer tests limit the test writer to those content areas that are capable of being stated in one or two words. And a multiple-choice question greatly reduces the opportunity for the students to guess the correct answer from one chance in two with a true-false test to *one in four or five, thereby increasing the reliability of the test.*

In preparing the objective multiple-choice test questions, teachers commonly find it helpful to keep the following suggestions in mind.

1. *Test students for important information and avoid trivia.* Teachers should resist the temptation to take the easy way out.

2. *Write the items clearly; avoid excessive verbiage-too many unnecessary difficult words; inappropriate choice of words;* and ***awkward sentence arrangement.*** Consider the following examples (the answer is option b.):

Poor: *The formulation of hypotheses*

a. is required to accomplish a descriptive study
b. guides the direction of research
c. states scientific fact
d. is proven correct by research

Better: *A hypothesis is a statement that a research*

a. employs as a technique for collecting data
b. uses as a guide in defining the nature of the study
c. accepts as a proposition of scientific facts
d. proves correct in the course of scientific investigation

3. Do not give the correct answer away with irrelevant clues. For example: in the following illustration the use of the indefinite article "an" gives the answer "electron" away since it begins with a vowel:

Poor: *A subatomic particle that has a negligible mass and carries a unit negative electrical charge is an*

a. proton
b. neutron
c. molecule
d. electron

4. Make each item independent of other items. Teachers should avoid writing items that are interrelated, like the following: (the answers to the two questions are respectively ***d.*** and ***c.***)

A type of radiation that travels at the speed of light is

a. a beta particle
b. an alpha particle
c. a cathode ray
d. a gamma ray

This type of radiation has the following charge

a. positive
b. negative
c. no charge
d. electric

5. Avoid the use of negative questions. More errors in interpretation are associated with a negative question. For example (the answer is option a.):

Poor: *The nucleus of the following element does not contain neutrons*

a. hydrogen
b. sodium
c. helium
d. neon

Better: *With the exception of the following element, the nuclei of all elements contain neutrons*

a. hydrogen
b. sodium
c. helium
d. neon

6. Avoid lifting a statement verbatim from a textbook or other sources. Verbatim statements are frequently ambiguous when ***they are used out of context.*** For example:

Poor: *Shale is clay that has become rock, mainly by pressure.*

Better: *Clay that has become rock through the action of the earth's pressure is termed shale.*

Conclusion

The English teacher-made tests should contain clear and concise directions. The students should be provided with a brief statement of the purpose of the test. Also, the students should be informed as to the length of time available for completing the test, the procedure for recording the answers, and how the test is to be scored. This information should be provided in the form of written directions.

With inexperienced students test-takers, it is advisable to provide practice-test items to verify and check that the directions are understood by the students test-takers.

Where appropriate, the students should be instructed as to what to do about cheating or misbehavior during their taking the tests.

The items comprising an objective test can be arranged in a manner to assist both the student

test taker and the teacher scorer. Measurement experts recommend that items be grouped together according format: short-answer, matching, true-false, and multiple-choice tests. Within each test type, those items dealing with the same subject matter can be placed together. This will present the students with orderly, integrated arrangement, rather than a disorganized mosaic of just so many bits and pieces of knowledge.

Finally, educators advise that the test items be arranged in order of their difficulty, from easy to hard. Most commonly, the test will begin with ***true-false items, followed in order by matching items questions, short-answer items, multiple-choice questions, and finally essay questions.***

References

1. Crooks, T. J. (2001). **The Validity of formative Assessments.** Paper for The Annual Conference of the British Educational Research Association. UK.
2. Ebel, R. L. and Frisbie, D. A. (1990). **Essentials of Educational Measurement.** (5th ed.) Englewood Cliffs, N.J.: Prentice-Hall.
3. Ericksen, S.C. (1995) **The Essence of Good Teaching.** San Francisco: Jossey-Bass.
4. Genesee, F. and Upshur, J. A. (2001). **Classroom-based Evaluation in Second Language Education.** USA: Cambridge University Press.
5. Gronlund, N. E. and Linn, R. (1990). **Measurement and Evaluation in Teaching.** (6th ed.) New York: Macmillan.
6. Jacobs, L. C. and Chase, C. I. (1992). **Developing and Using Tests Effectively: A Guide for Faculty.** San Francisco: Jossey-Bass .
7. McKeachie, W. J. and Svinicki, M. (2006). **Mckeachie's Teaching Tips** (12th ed.). Boston, MA.: Houghton Mittin .
8. Mehrens, W. A. and Lehmann, I. J. (1991). **Measurement and Evaluation in Education and Psychology.** (4th ed.) New York: Holt, Rinehart & Winston.
9. SEAMEO. (2003). **Language Testing.** Singapore: RELC.