JURNALHOLISTICS VOLUME 2 NOMOR 3 JUNI 2010 ISSN 2085-4021

THE RELIABILITY OF ESSAY TEST

Eriza
Sriwijaya State Polytechnic
Palembang
Phone: 0711-353414
Department of English

ABSTRACT

Besides validity and accuracy, a good test should also have reliability. It is an essential **factor of** an educational measurement. Generally it is difficult to obtain high reliability of **an essay** test. Essay test is often lack of accuracy and causes low reliability. Since it is not **easy to** have high reliability coefficient, The test item and the grading system must be **designed** in such a way in order to improve the reliability coefficient. This article will **look at** the causes of low reliability, how to estimate and improve the reliability of an essay **test.**

KEYWORDS

Essay test, reliability, test design, grading system

1. INTRODUCTION

At the end of semester or chapter a teacher may evaluate the student's achievement. The test result can represent evidence of learning outcomes. A test is a measuring instrument which provides information about students. It measures how well the students have mastered the material and determines how far the instructional objectives have been achieved. Gronlund (1976:6) describes " evaluation may be defined as a systematic process of determining the extent to which instructional objectives are achieved by the pupils". According to Kubiszyn and Jurs (1993:14) a test represents an attempt to provide objective data that can be used with subjective impressions to make better, more defensible decisions. The teacher gives a test to provide objective information , which is combined subjective, commonsense impressions to make better educational decisions.

Essay test item is the most popular form. This kind of test is usually used to measure higher-level learning outcome. Essay test item is relatively easy to prepare but difficult to grade. Factors such as test design, test administered, the difficulty of the test, grading criteria influence reliability score. According to Jacobs (1991:1) technically, reliability shows the extent to which test scores are free from errors of measurement. No classroom test is perfectly reliable because random errors operate to cause scores to vary or to be inconsistent, the goal is to minimize the inevitable errors of measurement and this can increase reliability. This paper discusses the reliability of essay test, how to estimate it and how to improve it.

2. CONCEPTUAL FRAMEWORK

2.1. The Essay Test.

Essay test item requires the students to create an explicit answer. The item is usually easy to design. This type is an effective measuring tool to investigate students' ability to think critically and originally. Cangelosi (1990) : 116) defines essay test item as a task to which the student is to respond by writing a literary composition of at least one paragraph but normally not more than several page. Wiresma and Jurs (1990: 69) point out that an essay test is usually one for which the student is required to structure a response. He or she selects ideas and then present them according to his or her own organization and wording. It means essay test relates to language skills in this case is writing skill. Essay test can be considered to be an approach to meaning writing performance

2.2. Types of Essay Test Items

Kubiszyn and Borich (1993:106) divide essay test items into two categories:
1. Extended-Response item
This item allows the students to determine the length of complexity of response. This type is most useful at the synthesis or evaluation level of the cognitive taxonomy.
2. Restricted –Response item
This item asks a specific problem for which the students must recall proper information, organize it in suitable manner, derive a defensible conclusion, and express it within the limits of the asking problem.

2.3. The Relation between Essay Test Items and Instructional Objectives

Essay test may be used to measure general or specific outcomes of instruction. Wiersma and Jurs (1990;73) mention that essay test is potential for measuring higher-level or complex learning outcomes. Essay item provides the students with an opportunity to organize, analyze and synthesize ideas. While Kubiszyn and borich (1993:108) describe that learning outcomes that can be measured by essay test item as follows:

- Analyze relationship
- Arrange items in sequence
- Compare positions
- State necessary assumptions
- Identify appropriate conclusions
- Explain cause-and-effect relations
- Formulate hypotheses
- Organize data to support a viewpoint
- Point out strengths and weaknesses
- Produce a solution for problem
- Integrate data from several sources
- Evaluate the quality or worth of an item
- Create an original solution, arrangement or procedure

2.4. The Reliability.

Reliability is an essential factor of an educational assessment. It refers to the consistency in measuring. A test is reliable if it consistently measures what it measures. Phelan and Wren (2005:1) mention reliability is the degree to which an assessment tool produces stable and consistent result. Moreover Lado (1961 In Ren 2007) claims the if test scores are stable, the test is reliable. If, however, the test scores fluctuate, the test is unreliable.
Reliability relates to getting the same score or close to it every time the teacher or lecture measures the students with the test. The test may not measure what it was intended but it is measuring in a consistent manner. The coefficient between the two test is a numerical value that must be calculated.

3. DISCUSSION

3.1 Factors Influence the Reliability of Essay Test

Factors influence the reliability of essay test can be varied, such as test design, the procedure or administration of the test, the difficulty of the test, grading system etc, even factors irrelevant to the item content can influence scores of essay test. Essay written in neat and good handwriting will typically be assigned higher scores than those written in poor handwriting.

Heat, light, noise, confusing direction and different testing time for different students are environmental factors which interfere the student performance and influence the test score.

Unstructured and unfocused questions can harm both the students and the raters. Look at the example : (Kubiszyn 1993:109)

Discuss the value of behavioral objectives.)

This question is unstructured and unfocused. It does not give the limit response, and does not establish a policy for grammar, spelling and punctuation. Different raters would probably have different idea and criteria for good and correct answer. It just depends on the raters. A lengthy answer with poor grammar and good content might get a high grade, a low grade, or an intermediate grade. Lets compare to the following example :

Behavioral objectives have enjoyed increased popularity in education over the past several years. In your text and in class the advantages and disadvantages of behavioral objectives have been discussed. Take a position for or against the use of behavioral objectives in education and support you position with at least three of the arguments covered in class or in the text. Although spelling, punctuation and grammar will not be counted in your grade, try to attend to these factors to the best of your ability.

Grading system is one factors influences the reliability of essay test. Grading the students response should base on criteria and methods. Without using grading criteria and applying good method, raters usually not consistent in their interpretation to the students' response and their impression of students influence their grading. A students score might be biased if his or her paper is graded immediately after an unexceptionally good or poor paper, and or read near the beginning or end of the grading session.

3.2 How to Estimate the Reliability of Essay Test

There are several types of reliability such as internal reliability, equivalent reliability, stability reliability, inter-rater reliability and intra-rater reliability. Essay test is usually assessed its inter-rater reliability and intra-rater reliability.

Inter-rater reliability is the degree of agreement among raters. It refers to the consistency of scores that are assigned by two raters. Intra-rater reliability represent the consistency of repeated measurements made by the same rater. Inter-rater and intra-rater reliabilities can be estimated through a number of statistics as following:

A. Pearson Product Moment Correlation Coefficient (PMCC)
The formula for calculating Pearson Product Moment Correlation Coefficient is:

$$r = \frac{\sum XY - \overline{XY}}{SDX * SDY}$$

(from How to Compute & Interpret Pearson 's Product Moment Correlation at http://www.slideshare.net/siddharth4mba/how-to-compute-interpret-...)

Example of calculating correlation coefficient using Pearson Product Moment formula

Data are from Popham (1973:91)

Subject	test 1 (X)	test2 (Y)	X*Y	X^2	Y^2	(X - Xr)
a	49	42	2058	2401	1764	7.9
b	46	42	1932	2116	1764	4.9
c	44	44	1936	1936	1936	2.9
d	44	40	1760	1936	1600	2.9
e	42	43	1806	1764	1849	0.9
f	40	38	1520	1600	1444	-1.1
g	38	39	1482	1444	1521	-3.1
h	38	40	1520	1444	1600	-3.1
i	36	29	1044	1296	841	-5.1
j	34	37	1258	1156	1369	-7.1
Σ	411	394	16316	17093	15688	

(X - Xr)^2	(Y - Yr)	(Y - Yr)^2	(X- Xr) / Sx	(Y- Yr) / Sy	
62.41	2.6	6.76	1.762532456	0.641242711	1.13021109
24.01	2.6	6.76	1.093216333	0.641242711	0.701017005
8.41	4.6	21.16	0.647005585	1.134506334	0.734031935
8.41	0.6	0.36	0.647005585	0.147979087	0.095743296
0.81	3.6	12.96	0.200794837	0.887874523	0.17828062
1.21	-1.4	1.96	-0.245415912	-0.345284537	0.084738319
9.61	-0.4	0.16	-0.69162666	-0.098652725	0.068230854
9.61	0.6	0.36	-0.69162666	0.147979087	-0.102346282
26.01	-10.4	108.16	-1.137837408	-2.564970843	2.918519776
50.41	-2.4	5.76	-1.584048156	-0.591916348	0.937624
200.9		164.4			6.746050613

X mean = 41.1
Y mean = 39.4
X^2 = 1709.3
Y^2 = 1568.8
XY = 1631.6

Sx = 4.482187
Sy = 4.054627

Rxy = 0.6746051

B. Spearman Correlation Coefficient.

The formula for calculating Spearman correlation coefficient is:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n-1)}$$

(from Spearman's rank correlation coefficient-Wikipedia,the free encyclo... at http://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coeffi...)

The example of calculating Spearman correlation coefficient. The data are from Popham (1973:91)

subject	measure1	Measure2	rank1	rank 2	d1	d_i^2
a	40	20	20	20	0	0
b	40	16	19	15	4	16
c	39	17	18	16	2	4
d	37	18	17	17	0	0
e	37	18	16	18	-2	4
f	36	20	15	19	-4	16
g	35	14	14	12	2	4
h	35	14	13	11	2	4
i	34	15	12	14	-2	4
j	34	12	11	8	3	9
k	33	13	10	9	1	1
l	33	15	9	13	-4	16
m	32	14	8	10	-2	4
n	32	11	7	5	2	4
o	30	12	6	7	-1	1
p	30	11	5	4	1	1
q	30	10	4	3	1	1
r	29	12	3	6	-3	9
s	28	9	2	1	1	1
t	26	10	1	2	-1	1
	20					100

ρ 0.924812

C. Cohen'Kappa Coefficient

Cohen 's kappa measures the agreement between two raters who each classify N items into C mutually exclusive categories. The formula for kappa is:

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

from Cohen's kappa-Wikipedia, the free encyclopedia at http://en.wikipedia.org/wiki/Cohen%27s_kappa)

$\Pr(a)$ is the relative observed agreement among raters and $\Pr(e)$ is the hypothetical probability of chance agreement, using the observed data to calculate the probabilities of each observer randomly saying each category. If the raters are in complete agreement then $\kappa = 1$. If there is no agreement among the raters then $\kappa \leq 0$.

Example of calculating kappa coefficient (taken from Cohen's kappa-Wikipedia, the free encyclopedia at http://en.wikipedia.org/wiki/Cohen%27s_kappa) as following: suppose that data related to people applying for a grant. Each grant proposal was read by two people and each reader either said "Yes" or "No" to the proposal. The data as follows, where rows are reader A and Columns are reader B:

	Yes	No
Yes	20	5
No	10	15

There were 20 proposals that were granted by both reader A and reader B, and 15 proposals that were rejected by both readers. Thus the observed percentage agreement is

$$\Pr(a) = \frac{20+15}{50} = 0.70$$

To calculate $\Pr(e)$ (the probability of random agreement) we note that:
Reader A said "yes' to 25 applicants and "No" to 25 applicants. Thus reader A said "Yes" 50% of the time. Reader B said " Yes" to 30 applicants and "No" to 20 applicants. Thus reader B said " Yes" 60% of the time. Therefore the probability that both of them would say "Yes" randomly is 0.50.x0.60= 0.30 and the probability that both of them would say "No" is0.50x 0.40=0.20. Thus the overall probability of random agreement is $\Pr(e) = 0.3 + 0.2 = 0.5$
The coefficient is :

$$\kappa = \frac{\Pr(a - \Pr(e))}{1 - \Pr(e)} = \frac{0.70 - .50}{1 - 0.50} = 0.40$$

3.3 How to Improve the Reliability of the Test

The reliability of an essay test relates to the item and grading system. It can be improved by increasing the test item design and the grading qualities. Poorly written question is one source of unreliable score. Question that does not specify response length could invite the students to give long answer. Long response usually needs time to read. It is difficult and tedious to score. In order to increase the reliability coefficient Kubiszyn and Borich (1990: 113) suggest to write good essay item, use several restricted- range item and use predetermined scoring system. Writing good items and using restricted range essay than extended range essay will help improve essay scoring reliability.

Using predetermined scoring scheme is one important point in measuring reliability of the test. The scheme contains criteria of scoring. If a teacher doesn't determine and specify the relevant criteria beforehand, the grading usually become inconsistent and bias. Moreover: the rules for constructing good essay test item are:

A. Use essay questions only to measure complex learning outcome.
If it measures comprehension and application achievement. Objective test will be appropriate. Essay test is appropriate when measured achievement is about giving reason, explaining relationship, describing data, and formulating conclusion.

B. Relate the question as directly as possible to the learning outcomes being measured.
Each question should be specifically designed to measure one or more well-defined outcomes. Relating the question to the learning outcomes being measured can help determine both content and form of the item.

The restricted response item is related quite easily to specific learning outcome because it so highly structured. The limited response expected from the students also make it possible to the test make to phrase the question so that its intent is communicated clearly to the students. The extended-response item however, requires greater freedom of response and typically involves a number of learning outcomes. This makes it more difficult to relate the question to the intended outcomes and to indicate the nature of the desired answer through the phrasing of the question. If the task is prescribed too rigidly in the question, the students' freedom to select, organize, and to present the answer is apt to be infringed upon. One practical solution is to indicate to the students the criteria to be used in evaluating the answer. For example, a parenthetical statement such as the following might be added: "Your answer will be evaluated in terms of its comprehensiveness, the relevance of its arguments, the appropriateness of its examples, and the skill with which it is organized." This clarifies the task to the students without limiting their freedom, and makes the item easier to relate to clearly defined learning outcomes.

C. Formulate questions that present a clear task to be performed.
Selecting precise term and carefully phrasing and rephrasing the questions with the desired response in mind will help clarify the task to the students. Avoid starting question with "who", " what", " when", "where", " name " and "list". These terms tend to limit the response to knowledge outcomes. Complex achievement is more appropriate by using word as " why", "describe", "explain", "compare", "relate", "contrast", "interpret", "analyze", "criticize" and " evaluate".

D. Don't permit a choice of question unless the learning outcome requires it.
If the students are permitted to choose, for example three out of five questions . The answer cannot be evaluated on a comparative basis. Giving students a choice among questions simply complicates the sampling problem further and introduces distortion into the test result.

E. Provide enough time for answering, and suggest a time limit in each question.
Essay questions are designed most frequently to measure intellectual skill and abilities therefore, time must be allowed for thinking as well as for writing. Thus, generous time limit should be provided. In measuring complex achievement it would seem better to use fewer questions in order to avoid measuring writing speed as achievement. (From Rules for Constructing Essay Test http://www. personal. psu.edu/faculty/s/r/sra113/602/essayexams.htm)

To avoid inconsistency, subjectivity and bias while grading or scoring, some rules can be applied as following:

A. Evaluate answer to essay in term of the learning outcome being measured.
Essay test is used to obtain evidence concerning the extent to which clearly defined learning outcome have been achieved. Thus, the desired student performance specified in these outcomes should serve as a guide both for constructing the questions and for evaluating the answers. If a question is designed to measure the ability to explain cause-effect relation, the answer should be evaluated in term of how adequately the students explain the particular cause-effect relation presented in the questions.

B. Score restricted –response answer by the point method using a model answer as a guide.
Scoring with the aid of a previously prepared scoring key is possible with the restricted-response item because of the limitation placed on the answer. The procedure involves writing a model answer to each question and determining the number of points to be assigned to it and to the parts within it. The distribution of points within an answer must take into account all scoreable units indicated in the learning outcomes being measured. For example , points may be assigned to the relevance of the examples used and to the organization of the answer, as well as to the content of the answer, if these are legitimate aspects of the learning outcome. It is usually desirable to make clear to the student at the time of testing the bases on which each answer will be judged (content, organization, and so on).

C. Grade extended-response answer by the rating defined criteria as a guide.
Extended-response items allow so much freedom in answering that the preparation of a model answer is frequently impossible. Thus, the test maker usually grades each answer by judging its quality in terms of a previously determined set of criteria, rather than scoring it point by point with a scoring key. The criteria for judging the quality of an answer are determined by the nature of the question and thus by the learning outcomes being measured. If students were asked to " describe a complete plan for preparing achievement test," for example, the criteria would include such thing as (1) the completeness of the plan (for example whether it included a statement of objectives, a set of specifications, and the appropriate types of items), (2) the clarity and accuracy with which each step was described, (3) the adequacy of the justification for each step, and (4) the degree to which the various parts of the plan were properly integrated.

Typically the criteria for valuating an answer are used to establish five levels of quality. Then as the answer to a question is read, it is assigned a letter grade or a number from one to five, which designates the readers' rating. One grade may be assigned on the basis of overall quality of the answer, or a separate judgment may be made on the basis of each criterion. The latter procedure provides the most useful information for diagnosing and improving learning and should used wherever possible.

More uniform standards of grading can usually be obtained by reading answers to each question twice. During the first reading the paper should be tentatively sorted into five piles, ranging from high to low in quality. The second reading can then serve the purpose of checking the uniformity of the answer in each pile and making any necessary shifts in rating

D. Evaluate all of the students' answer to question before proceeding to the next question. Scoring or grading essay tests question by question, rather than student by students, makes it possible to maintain a more uniform standard for judging the answer to each question. This procedure also help offset the hallo effect in grading. When all of the answer on one paper are read together, the grader's impression of the paper as a whole is apt to influence the grades he assigns to the individual answer. Grading question by question, of course, prevents the formation of this overall impression of a student's paper. Each answer is more appropriate to be judged on its own merits when it is read and compared with other answers to the same question, than when it is read and compared with other answers by the same student.

E. Evaluate answer to essay questions without knowing the identity of the writer. This is another attempt to control personal bias during scoring. Answer to essay questions should be evaluated in terms of what is written, not in terms what is known about the writers from other contacts with them. The best way to prevent our prior knowledge from biasing our judgment is to evaluate each answer without knowing the identity of the writer. This can be done by having the students write their names on the back of the paper or by using code numbers I place of names. (From Rules for Constructing Essay Test http://www. personal. psu.edu/faculty/s/r/sra113/602/essayexams.htm)

4. CONCLUSION

Reliability of essay test item is an essential factor of a test. It can be estimated through several procedures. The good quality of writing test items and the grading system will enhance the reliability coefficient. Essay test is more effective in measuring higher-level learning outcome. Essay test item should reflect learning outcome specified in the instructional objectives.

REFERENCES

Gangelosi, James S, 1990. *Designing Test for Evaluating Students Achievement*. N.Y.: Longman Inc.

_____Cohen's kappa Wikipedia (http://enwikipedia.org/wiki/Cohen%27_kappa) accessed on April 23.2010

Gronlund, Norman. E, 1976. Measurement & Evaluation in Teaching. N.Y.: Machmilan Publishing Co. Inc.

_____How To Compute &Interpret Pearson Product Moment Correlation Coefficient (http://www.slideshare.net/siddharth4mba/how_to_compute_interpret_...) accessed on April 25.2010

Jacobs, Lucy C. (1991) Test Reliability(http://www. Indiana. Edu/~ best/ test~ reliability.shtml) accesed on April 21.2010

Kubiszyn, Tom and Borich, Gary. (1993) *Educational Testing and Measurement Classroom Application and Practice*. New York: Harper Collins Publisher

Phelan, Colin. And Wren Julie. (2005-05) Exploring Reliability in Academic Assessment (http://www uni.edu/chfsoa/reliabilityandvalidity.htm) accessed on March 23.2010

Popham,W. James and Sitronik Kennenth. A. (1973) *Educational Statistics Use* and *Intrepretation.* U.S.A.: Harper & Collin Publisher

Sun, Guanxin The Reliability of Essay Marking in High Stakes Chinese second Language Examination 2007 (http://www. findarticle. com/p/ articles/ mi_ 6934/i_2_42 /ai_28469148/ ?t…) accessed on April 21. 2010

Rules For Constructing Essay Questions(http://www.personal .psu.edu/faculty/s/r/sral113/602/essayexams.htm) accessed on April 5 2010.

Spearman's Rank Correlation Coefficient Wikipedia(http://en.wikipedia.org/ wiki/Spearman% 27_rank_correlation_coeffi…) accessed on April 23 2010.

Wiersma, William and Jurs, Stephen G., 1990. *Educational Measurement and Testing.* USA: Allin and Bacon.

DEPARTEMEN PENDIDIKAN NASIONAL POLITEKNIK NEGERI SRIWIJAYA PERPUSTAKAAN