Reliability in the Research Data on Language Learning

by
Mohammad Adnan Latief
adnanlatiefs@yahoo.com

University of Pittsburgh
State University of Malang
2009

Reliability of Research Data on Language Learning

by Mohammad Adnan Latief

Abstract: This article discusses reliability, factors affecting the degree of reliability, and the technique of estimating the reliability of language skills assessment results. The term assessment is used here to refer to both conventional testing and alternative classsroom assessment. The meaning of reliability is discussed in contrast to the meaning of validity. While validity of language assessment results refers to the degree of correctness of representation of the language skill being assessed, reliability of language skill assessment results refers to the degree of preciseness of the reprersentation of the language skill being assessed. The farther the language skills assessment result deviates from the actual level of the skill being assessed, the bigger the error is, and the lower the reliability of the language skills assessment is. The low degree of reliablity is caused by the physical or emotional constraints of the learners being assessed, of the raters, of the instrument, and of the assessment administration process. Estimating reliability means collecting evidence of consistency.

Key Words: validity, reliability, error of assessment, correctness, preciseness.

When we are asking,"Do these speaking scores correctly represent the students' speaking skill?", we are questioning the degree of validity of those scores to represent the speaking skill. The scores for speaking skill obtained from a written test, as usually done in the Open University (Universitas terbuka=UT), for example, gives doubt to us whether those scores can really represent the speaking skill in question. Don't those scores better represent the knowledge of speaking rather than the skill of speaking? If, the speaking scores were obtained from an interview test in English, for example, we might not question the degree of the validity of those speaking scores to represent the speaking skill.

However, if we suspect that the interview test was done in a hurry, for example, too many students were interviewed in a short time and, therefore, the interviewers were too tired, then we are questioning the objectivity of the interview test result. Questioning the the objectivity of the interview test result means believing that there have been errors in the assessment, some scores are believed to be too high and some too low from their actual level of the speaking skill. Questioning the objectivity of the scores means believing that the reliability of the scores is low.

The old paper-and-pencil TOEFL gives an example of writing skill assessment with low validity. The "Written Expressions" Section, which takes the form of multiple choice, is meant to assess the quality of the writing skill of the examinees. Each of the stems is made of an expression containing a grammatical or lexical mistake to be identified by the examinees. Those who can recognize the mistakes will get high scores that represent high skill in writing. This is just the problem. The examinees are considered to have high writing skill because they can identify the mistakes in expressions provided in the test, not because they can show a piece of high quality writing. Questioning the correctness of these scores (from the objective type test) in representing the writing skill means believing that these scores have low validity in representing the writing skill. "Don't these scores better represent the knowledge of Grammar rather than the skill of writing?".

My experience in taking an Advanced Writing course in the English Language Teacher Training Program (ELTTP) of IKIP MALANG in 1977 gives an example of low reliability of the writing skill assessment result. A student in that writing class always got C or D for his writing assignment. Once, he tried to cheat the writing instructor. He asked a friend whose writing skill was very good (his writing assignment always got A) to write an essy for him. Unfortunately, when the essay written by his friend was submitted, the essay (which was suppoesd to deserve A) still got C. This judgment of the quality of the essay by the instructor was based on who submitted the essay, not based on the actual quality of the essay.

This article tries to discuss the term reliability in contrast to validity of the results of language skills assessment. Since reliability and validity are the characteristics of scores or the results of some assessment (not the characteristics of an assessment instrument), the discussion here is related to the scores or the results of assessment of language skills. The term assessment is used here to refer to both conventional testing and alternative classroom assessment (Stiggins, 1994: 8). Conventional testing refers to standard types of evaluation, like objective types (True-False, Multiple Choice, Matching, Short Answer), and subjective types (essay, demonstration), administered at a certain time (in the middle and at the end of a quarter or a semester), to evaluate the achievement in learning or the proficiency, for the purpose of judging the quality of test takers' skills

level, or for the purpose of making decision for the test takers (giving scores, regrouping, passing/failing, keeping them at the same grade or promoting them to the next grade).

Alternative classroom assessment, on the other hand, refers to any activity (other than that in conventional testing) that involves systematic collection of information about the language skill the test takers are learning/acquiring, administered throughout the whole process of learning, to monitor the progress of learning, for the purpose of facilitating instructors in their giving the learners maximum help in learning (Hoy & Greg, 1994:4). The alternative classroom assessment for the learners' progress in learning writing, for example, may take the form of (the instructor's) examining and recording the learners' journals, their notebooks, their writings in wall magazines, their letters to their classmates or to the instructor. To assess the learnres' progress in learning speaking, the instructor may observe and record the learners' speaking activities while communicating to the instructor, to their classmates during the process of learning other subjects, during the breaks, while performing poetry reading, while acting in a play, etc. To assess the students' progress in learning vocabulary, the instructor may watch the choice of words the students are using while they are speaking or in the sentences they write for journals, notebooks, letters, etc. To assess the students' progress in learning grammar, the instructor may pay attention to the structure of sentences the students are using while they are speaking or the sentences they write for journals, notebooks, letters, etc.

DEFINING RELIABILITY

While validity refers to the degree of **correctness** of the writing skill assessment result in representing the writing skill being assessed (to what extent the result of a language skill assessment result doesn't mistakenly represent another language skill, or to what extent the result of speaking skill assessment result doesn't mistakenly represent the knowledge of speaking), reliability of the result of language skill assessment refers to the **preciseness** of the language skill assessment result in representing the actual level of the skill of the examinees. The result of a language skill assessment has high reliability if the result precisely represents (is very closed to, or is not too far away from, or gives good estimate of, or does not overestimate or underestimate) the true level of the skill being assessed. In other words, if the language skill assessment result is too far away

different from the true level of the skill being assessed, then the assessment result has low reliability. The ditance between the true level of the skill and the assessment result, then, determines the degree of reliability; the bigger the distance is between the language skill assessment result and the actual level of the skill being assessed, the lower the reliability of that assessment result is. The distance between the language skill assessment result and the real level of the skill being assessed represents errors of the assessment result. In other words, the bigger the errors in the assessment result are, the bigger the distance is between the assessment result and the actual level of the skill being assessed, and the lower the reliability of that assesment is.

Mathematically, the relationship between the language skill assessment result (X), the true level of the skill being assessed (T), and the errors (E) can be formulated as follows

$$\mathbf{X = T + E}$$

(Allen & Yen, 1979: 57, Ebel & Frisbie, 1985: 72)

The formula explains that (every) language skill assessment result (X) contains the mixture of the true level of the language skill being assessed (T) and the error (E). The amount of error (E) determines the degree of the reliability of the language skill assessment result (X). The bigger the error (E) is, the lower the reliability of the language skill assessment result (X) is, and similarly, the smaller the error (E) is, the higher the reliability of the assessment result (X) is. See Allen & Yen as quoted below.

> "As reliability of a test increases, the error score variance becomes relatively smaller. When error variance is relatively slight, an examinee's observed score is very close to his or her true score. However, when error variance is relatively large, observed scores give poor estimates of true scores" (Allen & Yen, 1979:73).

Some language testing experts define reliability as referring to consistency of the scores resulted from the assessment (See Djiwandono, 1996: 98, Gronlund, 1985: 86 for examples). Consistency is an important indicator for reliability, meaning that if an assessment result is (or the test scores are) consistent from one assessment to another, then the assessment result has (or the test scores have) high reliability. However, consistency is not the meaning of reliability, it is only an indicator of reliability. The

meaning of reliability (of a language skill assessment result) is preciseness (of the assessment result or the closeness of the X to T).

FACTORS AFFECTING THE DEGREE OF RELIABILITY

The main factor affecting the validity of language skill assessment result is the appropriateness of the procedure of the assessment (the appropriateness of the choice of instrument). An assessment of speaking skill using a paper-and-pencil test that requires the examinees to show their speaking skill by writing and based on the writing the speaking skill is estimated, for example, will result in the speaking scores with low validity (weak construct-validity evidence), which means that the speaking scores better represent the *knowledge* of speaking rather than the *skill* of speaking. If the assessment of speaking skill is administered using an interview test, which requires the examinees to show the skill of speaking by actually talking and based on the talking activity the speaking skill is estimated, then the result of the speaking assessment (or the scores) will have higher validity (with higher construct validity evidence).

> "When we interpret test scores as a measure of a particular construct, we are implying that there is such a construct, that it differs from other constructs, and that the test scores provide a measure of the construct that is litle influenced by extraneous factors. Verifying such implications is the task of construct validation." (Gronlund, 1985: 72).

However, if the interview test only requires the examinees to anwer the questions with yes-or-no answers or only oral short answers, then the result of the speaking assessment (or the scores) will still have low validity (with low content validity evidence).

> "Content validation is the process of determining the extent to which a set of test tasks provides a relevant and representative sample of the domain of tasks under considerations." (Gronlund, 1985:59).

While low validity of assessment result means that the scores resulted wrongly represent another skill than the skill being assessed, low reliability of assessment result means that the scores resulted contain big errors and so give poor estimates for

(overestimate or underestimate) the true level of the skill being assessed. The poor estimates of the assessment result may be caused by (1) the inability of the examinees to show the best performance, (2) the inability of the instrument to solicit the best performance from the examinees, (3) the inability of the raters to give objective judgement about the level of the skill being assessed. Ebel & Frisbie (1985: 73) said :"Reliability depends on the nature of the group tested, the test content, and the conditions of testing."

Not the Examinees' Best Performance.

Errors in assessment that cause the scores to underestimate the true level of the skill being assessed may happen because the examinees are not in their best condition when the assessment is being administered due to the phyisical as well as emotional constraints. They may be sick, tired, hungry, emotionally disturbed, not concentrating, sleepy while the assessment is conducted. Doping or any situation that makes the examinees over-active or exessively happy, on the other hand, can also cause the assessment to result in scores which overestimate the actual level of the skill being ssessed. To avoid the errors of either underestimating or overestimating the true level of the skill being assessed, therefore, the assessment should be conducted in such situation that the constraints can be minimized. The assessor should select the best conducive atmosohere to make sure that the examinees are not having those constraints while the assessment is being administered.

Not the Raters' Most Objective Judgment.

Like the errors coming from the examinees' physical as well as emotional constraints, errors in assessment that cause the scores to underestimate the true level of the skill being assessed may happen because the raters who give the judgment to the quality of the skill being assessed are not in their most natural and objective physical as well as emotional mode. They may be sick, tired, hungry, emotionally disturbed, not concentrating, sleepy while giving judgement. Doping or any situation that makes the raters over-active or exessively happy, on the other hand, can also cause the judgement to result in scores which overestimate the actual level of the skill being ssessed. To avoid the errors of either underestimating or overestimating the true level of the skill being

assessed, therefore, the judgement process should be conducted in such situation that the constraints can be minimized. The raters should select the best conducive atmosphere to make sure that they are not having those constraints while giving the judgement.

The Assessment Instrument Being too Short.

An assessment for knowledge of English Grammar which asks 100 questions will result in the assessment scores with higher reliability than the same assessment which asks only 25 questions. Similarly, an interview to assess the speaking skill which allows 30 minutes for each examinee to talk will result in the scores for speaking skill with higher reliability than the same interview which allows only 5 minutes for each examinee to talk. In short, an assessment which asks more questions or which allows more time for the examinees to show their performance will result in scores with higher reliability than the same assessment which asks fewer questions and allows shorter time for the examinees to show their performance. (Ebel & Frisbie, 1985: 84)..

The Assessment Instrument Content being Heterogenous

An assessment for written integrated English skills which covers the skill of Reading, the skill of Writing, the knowledge of Grammar, and the knowledge of Vocabulary will result in the scores with lower reliability than an assessment for only one specific language skill or the knowledge of language component, like an assessment for the skill of Reading, the skill of Writing, the knowledge of Grammar, or the knowledge of Vocabulary. In other words, an instrument with a number of questions designed to assess many different language skills or knowledge of language components (heterogeneous) will result in the scores with lower reliability than an instrument with the same number of questions designed to assess one specific language skill or one specific knowledge of language component (homogeneous). (Ebel & Frisbie, 1985: 84).

The Assessment Questions Being too Esay or too Difficult.

An assessment for Listening Comprehension skill, for example, which asks so difficult questions that only 10 percent of the examinees can answer all the questions correctly, or so easy questions that almost all of the examinees can answer all the questions correctly will result in scores with lower reliability than the same assessment

which asks questions with moderate difficulty that about 35 to 85 percent of the examinees can anwer all the questions correctly.(Ebel & Frisbie, 1985: 85). So, the level of difficulty of the questions in the assessment instrument influences the degree of reliability.

The Type and Quality of Assessment Instrument

An instrument designed to assess the knowledge of Vocabulary which contains 100 multiple-choice-type questions could produce scores with the same degree of reliability as the same assessment instrument which contains 150 True-False-Type questions. Or, a multiple-choice-type instrument with more plausible distracters can produce scores with higher reliability than the same instrument with less plausible distracters. So, the type of instrument and the quality of the distractors influence the degree of reliability. (Ebel & Frisbie, 1985: 85).

Cheating in the Assessment

If the examinees are not strictly watched during the assessment process, they might copy each other's answers or they might copy their notes they have prepared. If this cheating happens then the assessment will result in low reliability scores So, honesty of the examinees in answering the assessment questions affects the degree of reliability.

Uncomfortable Place and Time of Assessment

An assessment conducted in an uncomfortable room; too hot, too cold, too windy, too crowded, too small, or too noisy, and in uncomfortable time; at 2.00 p.m. after the examinees have worked all morning, for example, will result in the assessment scores with low reliability.

ESTIMATING THE DEGREE OF RELIABILITY

Estimating the degree of reliability of scores or the results of language skill assessment means verifying or confirming whether the scores or the results of the assessment have high degree of preciseness. The evidence assumed to indicate high degree of preciseness of scores or the results of language skill assessment is the consistency of the scores. It is assumed that if the scores or the result of assessment

precisely represent the actual level of the skill being assessed or if the scores do not contain too big mistakes (if the distance between the scores and the actual level of the skill being assessed is not too big), then the scores or the result of the assessment will be consistent. Conversely, if the scores or the results of the assessment do not precisely represent the actual level of the skill being assessed or if the scores contain too big mistakes (if the distance between the scores and the actual level of the skill being assessed is too big), then the scores or the result of assessment will not be consistent.

Estimating the degree of reliability, therefore, refers to an effort to collect evidence of consistency to verify or to confirm the reliability. In other words, if the scores or the results of assessment are provided with evidence of high consistency, then the scores or the results of the assessment convincingly have high degree of reliability/ preciseness. For the scores or the results of conventional testing (from one single time /shot testing), the evidence of consistency is highly needed. This is because the single time /shot testing potentially suffer from several problems (See factors affecting the degree of reliability).

The evidence of consistency may be collected by computing the correlational index of the two sets of scores from two times testing (using test-retest method), the scores from one test with the scores from its parallel form (using equivalent form method), by computing split-halve correlation, by computing internal consistency using Statistics formula; KR-20 or KR-21 and by cross-checkiong inter raters' agreement. (See Djiwandono, 1996:99 and Ebel & Frisbie, 1986: 75 for more details).

For the results of alternative classroom assessment, which are relatively safer from the factors affecting the degree of reliability, the evidence of high consistency is already provided through the process of the assessment. The result of assessment is not based on one single time /shot assessment, it is instead based on several times assessment throughout the whole process of instruction, from the beginning minutes to the last minutes of instruction, from the first instruction to the last instruction in the quarter or the semester.

FINAL REMARKS ON RELIABILITY

Language skills assessment should be planned and conducted in the best possible way to get the judgement of the skills being assessed that **correcly** represents the skills

being assessed or does not wrongly represent another skill not being assessed (have high construct and content validity) and **precisely** represents the level of the skills being assessed, or doesn't too far overestimate or underestimate the true skill being assessed (have high reliability). Low reliability of assessment results indicates that there have been big errors in the process of assessment. The errors may happen because the examinees cannot show their best performance during the assessment process, the raters cannot give their most objective judgement, the assessment process is not conducive enough for the examinees to show their best performance, the content of the assessment instrument is too heterogeneous, the questions are too difficult or too easy, or the distractors are not quite plausible.

The evidence of high consistency of language skills assessment result interpreted from conventional testing can be collected by computing the correlational index from two sets of scores obtained from the method of retesting with the same or parallel form instruments, from split-halves scores, or by computing the internal consistency, and by crosschecking the inter raters' judgment agreement. For the results of language skills assessment interpreted from altentaive classroom assessment, the evidence of high consistency is already provided through the prolonged process of its assessment.

REFRENCES

Allen, Mary, J., Yen, Wendy, M., 1979. *Intouduction to Measurement Theory*. Monterey, California: Brooks/Cole Publishing Company.

Djiwandono, M. Soenardi, 1996. *Tes Bahasa dalam Pengajaran*. Bandung: Penerbit ITB.

Ebel, Robert, L., Frisbie, David, A. 1986. *Essentials of Educational Measurement*. Englewood Cliffs, New Jersey: Prentice Hall, Inc.

Gronlund, Norman, E., 1985. *Measurement and Evaluation in Teaching*. New York:Macmillan Publishing Company.

Hoy, Cheri., Greg, Noel. 1994 *Assessment. The Special Educator's Role*. Bemont, California: Brooks/Cole Publishing Company.

Stiggins, Richard, J. 1994 *Student-Centered Classroom Assessment*. New York: Macmillan College Publishing Company.