CRITERION-REFERENCED TESTING AND ITS PROBLEMATIC ESTIMATE OF RELIABILITY

Khairuddin
Department of Medical Record and Health Information
State Health Polytechnic of Malang

Abstract: Two different families of test criterion-referenced and norm-referenced tests have been around for use of decision-making bases in education. One kind can be better explained by comparison with the other. The article does so here and there along the way in its attempts of presenting and discussing about criterion-referenced test in language testing, especially the contrastive practices within this test family as opposed to the other kind, which has so close a relationship to classroom teachers on a daily basis. The article describes its purpose, hence several kinds of CRT as a result of different purposes of testing, what is measured and how measuring is done, how is test results interpreted, and how reliability as one of the quality of good testing is estimated. Discussion of validity as another test quality indicator is purposely put aside here since in this area the contrastive practice is only slimly found.
Keywords: *double entry diaries, strategy, reading comprehension.*

Criterion-referenced testing (CRT) is a tool or instrument functions as a test which measures a student's performance according to a particular standard or criterion which has been agreed upon even before classroom instruction is started (Richards, Platt, and Weber, 1985; Cohen, 1994; Djiwandono, 2007). CRT is usually produced to measure well-defined and fairly specific instructional objectives (Brown, 2005), often are specific to a particular course, program, school district, or state. However, objectives come in many forms. Other objectives might be defined in term of tasks we would expect the students to be able to able to perform by the end of the term, or experiences we would expect them to go through. For example, by the end of the term the students will watch at least five English language movies with no subtitles.

In addition to measuring the amount of material learned as representing the achievement of the objectives of instruction by the students, CRT also attempts to score the test and report the results to the students in the form of percentages of questions students answered correctly. These percentage scores can then be directly related to the material taught in the class and related to a previously established criterion level for passing the test.

CRT be also designed to give test-takers feedback, usually in the form of grades, on specific course or lesson objectives (Brown, 2002). Classroom tests involve the students in one class, and is connected to a curriculum, thus the result of the tests are expected to be useful for the pursuit of teaching effectiveness in the class and the curriculum repair efforts, or what Oller (1979, P. 52) called "instructional value." In a criterion-referenced test, the distribution of students' scores across a continuum may be of little concern as long as the instrument assesses objectives. From the results of CRT several decision-makings like the following can be based.

Decision-making based on CRT use

CRt results are usually used to base the making of classroom-related decision. Subject to the purposes of conducting the CRT, two types of decisions are usually made out of the application of CRT: classroom-level achievement decisions and classroom-level diagnostic decisions.

Classroom-level achievement decisions are decisions about the amount of learning that students have accomplished. Such tests are typically administered at the end of the term, and such decisions may take the form of deciding which students will be advanced to the next level of study, determining which students should graduate, or simply for grading the students. Teachers may find themselves wanting to make rational decisions that will help improve their students' achievement. Or they may need to make and justify changes in curriculum design, staffing, facilities, materials, equipment, and so on. Such decisions should most often be made with the help of achievement test scores.

Making decisions about the achievement of students and about ways to improve their achievement will at least partly involve testing to find out how much each person has learned within the program. Thus, achievement tests should be designed with very specific reference to a particular course. This link with a specific course usually means that the

achievement tests will be directly based on course objectives and will therefore be criterion-referenced. Such tests will typically be administered at the end of a course to determine how effectively students have mastered the instructional objectives.

Achievement tests must not only be very specifically designed to measure the objectives of a given course, but also must be flexible enough to help teachers readily respond to what they learn from the tests about the students' abilities, the students' needs, and the students' learning of the course objectives. In other words, a good achievement test can tell teachers a great deal about their students' achievement and about the adequacy of the course. Hence, while achievement tests should definitely be used to make decisions about students' levels of learning, they can also be used to affect curriculum changes.

From time to time, teachers may also take an interest in assessing the strengths and weaknesses of each individual student in terms of the instructional objectives for the purpose of correcting an individual's deficiencies "before it is too late." To that end, classroom-level diagnostic decisions are typically made at the beginning or middle of the term and are aimed at fostering achievement by promoting strengths and eliminating the weaknesses of individual students. Naturally, the primary concern of the teacher must be the entire student. Clearly, this last category of decision is concerned with diagnosing problems that students may be having in leaning process. While diagnostic decisions are definitely related to achievement, diagnostic testing often requires more detailed information about which specific objectives students can already do well and which they still need to work on. The purpose is to help students and their teachers to focus their efforts where they will be most effective.

As with achievement tests, diagnostic tests are designed to determine the degree to which the specific instructional objectives of the course have already been accomplished. Hence, they should be criterion-referenced in nature. While achievement decisions are usually focused on the degree to which the objectives have been accomplished at the end of the program or course, diagnostic decisions are normally made along the way as the students are learning the language. As a result, diagnostic tests are typically administered at the beginning or in the middle of a language course. In fact, if well constructed to reflect the instructional of objectives, one CRT in three equivalent forms could serve as a diagnostic tool at the beginning and midpoints in a course as an achievement test at the end.

Perhaps the most effective use of a diagnostic test is to report the performance level on each objective (in a percentage) to each student so that they can decide how and where to most profitably invest their time and energy. For example, telling the student that she scored 100% on the first objective (selecting the main idea of a paragraph) but only 20% of the second objective (guessing vocabulary from context) would tell that student that she is good at finding the main idea of a paragraph but needs to focus her energy on guessing vocabulary from context.

It would also be useful to report the average performance level for each class on each objective (in percentage terms) to the teacher(s) along with indications of which students have particular strengths or weaknesses on each objective.

Interpretation of CRT

The interpretation of scores on a CRT is considered absolute. Each student's score is meaningful without reference to the other student's scores, like in NRT interpretation. In other words, a student's score in a particular objective indicates the per cent of the knowledge or skill in that objective that the student has learned. Moreover, the distribution of scores on a CRT need not necessarily be normal. If all the students know 100% of the material on all the objectives, then all the students should receive the same score with no variation at all. The purpose of a CRT is to measure the amount of learning that a student has accomplished on each objective. In most cases, the students should know in advance what type of questions, tasks, and content to expect for each objective because the question content should be implied (if not explicitly stated) in the objectives of the course.

In terms of the type of interpretation, each student's performance on a CRT is compared to a particular criterion in absolute terms. Some confusion has developed over the years about what *criterion* in a criterion-referenced testing refers to. This confusion is understandable because two definitions have evolved for criterion. For some authors, the material that the students are supposed to learn in a particular course is the criterion against which they are being measured. For other authors, the term criterion refers to the standard, also called a criterion level or cut-point against which each student's performance is judged. For instance, if the cut-point for passing a CRT is set at 70 per cent, that is the criterion level.

Related to the criterion in CRT, criteria to which the students' performance are referred to, are not firstly and merely a certain number of scores (for example 24 correct answers of 30 test items) or a percentage of achievement (for example 70% or 80% of the whole test). Nor are the referenced criteria merely a certain score as the passing score or cut-off point decided alone by a test developer (Djiwandono, 2007). Basing achievement decisions on minimum score or percentage of correct answers may result in the unclearness of the concept of criteria itself in the application of CRT. The aforementioned numbers or points concerning with minimum scores and percentage of correct answers, if ever used, should only be the byproduct of a very clearly specified and formulated set of criteria, which have clear, easy and open outlook. The number or points should be used in close relation to the discretely specified kinds and levels of ability required to pass the test. The specified aspects of the skill or ability further should be translated to detailed indicators of the mastery of the ability which can be searched, observed, and verified anytime.

For example, in a paper-writing test, the minimum criteria considered adequate that the students' performance should show would cover discreet specific aspects which as a whole can represent as an adequate skill in paper writing that may include content, organization, grammar, vocabulary, and writing technicality, like exampled by Djiwandono (1989) as follows:

Specific Criteria of Test of Writing Skill

No	Aspects of Writing Skills	Indicator of Minimum Competence Achieved
1.	CONTENT	Discussion content is relevant to topic topic of discussion is mastered Discussion coverage is adequate
2.	ORGANIZATION	Discussion is developed based on main ideas topic sentences Topic sentences are well-organized Topic sentences are well-developed
3.	GRAMMAR	Sentences are built grammatically Sentences are used effectively Phrases and words are formed grammatically
4.	VOCABULARY	Vocabulary size is adequate Word choices and uses are appropriate
5.	SPELLING AND TECHNICALITY OF WRITING	Spelling and punctuation are rule-appropriate

As an example of CRT application, the detailed specified criteria in the example contain indicators of minimum ability that test takers should achieve. A person can pass the paper writing test only when the minimum level of ability can be identified in the test-taker's writing production, to be considered as having the required minimum ability and therefore is entitled to a minimum score 70 or 75 or a passing grade C or B indicating that the test-takers have shown the minimum required ability. Meanwhile, the performances which cannot reach the minimum level of, say, c or B are given lower grade D. The same thing, performances which reach higher than the minimum level of ability prescribed can be given higher grade A.

Another way to set up the criteria as judgment basis is by using the performance of a criteria group. The criteria group is a group of people whose expertise is a public knowledge or they are acknowledged as having accountable ability in the field targeted in the test. Their performance in doing the task as required by the test is used as criteria or reference in deciding the level of ability expected to be achieved by test takers in such a test.

As an example that Djiwandono (2007) gives regarding criteria group reference is the level of ability/skill of writing in Bahasa Indonesia by graduate students which are based on or referred to the performance of criteria group comprising of 10 senior Bahasa Indonesia and English professors. Based on their performances on a daily basis as educated people, senior language professors, and highly competent Bahasa Indonesia users, their performances can be accountably considered to represent and be used as necessary criteria to be referred to when judging other test takers ability in the relevant test objective. The argumentative writing assigned to the members of criteria group is judged based on the profiles of writing product developed in his research, presenting scores that range from 38 to 100. The scoring of writing performance by members of the criteria group found score 95 as the highest score and 79nas the lowest. By referring back to the profiles of good writing piece, 4 levels of writing ability are decided: A (Very Good), B (Good), C (Fair), and D (Bad), with score ranges of 90-100, 72-89, 57-71, and 34-56, respectively. Based on that, criteria of minimum level ability is decided at "fair" qualification at the very least, that range between scores 57-71 or grade C. The scores that fall below the minimum level of ability thus bear inferior grade D which means failing the minimum criteria required to pass the writing test.

Type of measurement

According to Brown (2005), with regard to type of measurement, NRTs are typically most suitable for measuring general abilities. In contrast, he addresses that CRTs are better suited to providing precise information about each individual's performance on well-defined learning points. For instance, if a language course focuses on a structural syllabus, the CRT for that course might contain four subtests on: subject pronouns, the a/an distinctions, the third person –s, and the use of present-tense copula. However, CRTs are not limited to grammar points. Subtest on a CRT for a notional-functional language course might consist of a short interview where ratings are made of the student's abilities to: perform greetings, agree or disagree, express an opinion, and a conversation. The variety and types of test questions used on a CRT are limited only by the imagination of the test developer(s).

Distribution of scores

Since NRTs must be constructed to spread students out along a continuum or distribution of scores, the manner in which test questions for an NRT are generated, analyzed, selected, and refined will usually lead to a test that produces scores which fall into a normal distribution (Brown, 2005; Djiwandono, 2007). In contrast, on a criterion-referenced final examination, students who have learned all the course material should all be able to score 100 per cent on the final examination. Thus, very homogeneous scores can occur on a CRT. In other words, very similar scores among students on a CRT may be perfectly logical, acceptable, and even desirable if the test is administered at the end of a course. In this situation, a normal distribution of scores may not appear. In fact, a normal distribution on CRT scores may even be a sign that something is wrong with the test, with the curriculum, or with the teaching (Brown 2004).

Test structure

Differences also arise in the test structure for the two families of tests. Typically, an NRT is relatively long and contains a wide variety of question content types. Indeed, the content can be so diverse that students find it difficult to know exactly what is being tested. Such a test is usually made up of a few subtests on rather general language skills. In contrast, CRTs usually consist of numerous shorter subtests. Each subtest will typically represent a different instructional objective. If a course has twelve instructional objectives, the associated CRT will usually have twelve subtests. Sometimes, in courses with many objectives, for reasons of practicality, only a sub-sample of the objectives will be tested. For example, in a course with 30 objectives, it might be necessary due to time constraints to randomly select 15 of the objectives for testing, or to pick the 15 most important objectives (as judged by the teachers). Because of the number of subtests involved in most CRTs, the subtests are usually kept short.

For reasons of economy of time and effort, the subtests on a CRT will sometimes be collapsed together, which makes it difficult for an outsider to identify the subtests. For example, on a reading comprehension test, the students might be required to read five passages and answer four multiple-choice questions on each passage. If on each passage there is one fact question, one vocabulary question, one cohesive devise question, and one inference question, the teachers will most likely consider the five fact questions (across the five passages) together as one subtest, the five vocabulary questions together as another subtest, the five cohesive device questions together as yet another subtest, and the five inference questions together as the last subtest. In other words, the teachers will be focusing on the question types as subtests, not the passages, and this fact might not be obvious to an outsider observer.

Finally, the two families of tests differ in the knowledge of the questions that students are expected to have. Students rarely know in any detail what content to expect on an NRT. On a CRT, good teaching practice is more likely to lead to a situation in which the students can predict not only the questions formats on the test, but also the language points that will be tested. If the instructional objectives for a course are clearly stated, if the students are given those objectives, if the objectives are addressed by the teacher, and if the language points involved are adequately practiced and learned, then the students should know exactly what to expect on the test, unless for some reason the criterion-referenced test is not properly referenced to the criteria (i.e., the instructional objectives).

This can often lead to complaints that the development of CRts will cause teachers to "teach to the test" to the exclusion of other more important ways of spending classroom time. Not all elements of the teaching and learning process can be tested; teaching to the test should nevertheless be a major part of what teachers do. If the objectives of a language course are worthwhile and have been properly constructed to reflect the needs of the students, then tests based on those objectives should reflect the important language points that are being taught. Teaching to such a test should help teachers and

students stay on track, and the test results should provide useful feedback to both groups on the effectiveness of the teaching and learning [processes. In short, teaching to the test, if the test is a well developed CRT, should help the teachers and the students rather than constrain them.

Reliability of CRT

As noted previously, CRTs will not necessarily produce normal distributions, especially if they are functioning correctly. If all the students have learned the material, the tester would like them all to score 100 per cent on the end-of-course achievement CRT. Hence, a CRT that produces little variance in scores is an ideal that test developers seek in developing CRTs. In other words, a low standard deviation on the course post-test may actually be a positive byproduct of developing a sound CRT. This is quite the opposite of the goals and results when developing a good NRT, which ideally should approximate a normal distribution of scores to the greatest extent possible (Brown 2005).

As far back as Popham and Husek (1969) the appropriateness of using correlational strategies for estimating the reliability of CRTs was questioned, because such analyses all depend in one way or another on normal distribution and a large standard deviation. Consider the test-retest and equivalent-forms strategies. In both cases, a correlation coefficient is calculated. Since correlation coefficient are designed to estimate the degree to which to sets of numbers go together, scores that are tightly grouped, because of skewing or homogeneity of ability levels, will probably not line the students up in similar ways. As that standards deviation approaches zero, so do any associated correlation coefficients. Correlation coefficients used for estimating interrater and intrarater reliability will be similarly affected by such circumstances. Therefore, reliability measure strategies in NRT may be quite inappropriate for CRTs because CRTs are not developed for the purpose of producing variance in scores.

The following example split-half reliability case would show how the application of NRT-based reliability estimate strategy is inappropriate when applied to measure the reliability of a CRT.

Odd-numbered: 13, 13, 12, 12, 12, 11, 12, 11, 12, 10, 9, 10, 11, 11, 9, 9, 8, 8, 9, 9, , 5, 8, 8, 6, 6, 9, 7, 6, 3, 5

Even-numbered: 14, 14, 14, 12, 12, 10, 9, 9, 7, 8, 9, 8, 7, 7, 8, 8, 8, 8, 7, 6, 9, 6, 6, 7, 7, 3, 5, 5, 7, 3

Number of students: 30
Number of items in each group: 15
Mean score of odd-numbered items: 9.2
Standard deviation: 2.6
Mean score of even-numbered items: 8.1
Standard deviation: 2.8
Correlation coefficient split-half reliability: 0.64
(Pearson-product moment)
Full-test reliability: 0.80
(Cronbach alpha)

Internal consistency reliability estimates the consistency of a test using only information internal to a test, that is available on one administration of a single test. Split-half method is one way to estimate the consistency or reliability of an NRT test. The procedure requires a more or less normal distribution of scores to the greatest extent possible, hence a large standard deviation. As the procedure, a single test is divided into two parts on the basis of odd-numbered and even-numbered items. The odd-numbered and even-numbered items are scored separately as though they were two different forms. A correlation coefficient is then calculated for the two sets of scores. This coefficient gives the reliability for either the odd-numbered items or the even-numbered items—either half, but just half of the test. The adjustment of the half-test correlation to estimate the full-test reliability is accomplished by using the Cronbach Alpha coefficient, one of the others.

The mean of the 30 students' NRT-based scores in the *odd* category is 9.20, and in the *even* category 8.10. The standard deviation for either group respectively is 2.66 and 2.80. The split-half reliability is 0.66, and when converted to full-test reliability using Cronbach alpha coefficient, a reliability level of 0.80 is achieved. 0.80 coefficient point is a high reliability level given the highest point of correlation coefficient is 1. So interpretation can be made, that the test developed, by moving the decimal two places to the right, is 80 per cent consistent or reliable with 20 per cent measurement error.

Now look when the same procedure is applied to CRT-based scores, where distribution of the scores is not normal but in fact clustered around one extreme, to estimate the internal consistency of the test through split-half method.

Odd-numbered items: 13, 13, 13, 12, 12, 13, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 13, 13, 12, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 12

Even-numbered items: 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 13, 13, 13, 12, 12, 12, 12, 12, 14, 14, 13,

13, 13, 13, 13, 13, 13, 13, 13, 12

Number of students: 30
Number of items in each group: 15
Mean score of odd-numbered items: 12.56
Standard deviation: 0.50
Mean score of even-numbered items: 12.53
Standard diation: 0.56
Correlation coefficient split-half reliability: 0.42 (Pearson-product moment)
Full-test reliability: 0.60 (Cronbach alpha)

The mean of the 30 students' CRT-based scores in the odd-numbered category is 12.56, and in the even category 12.53. Looking at the superficial odd-numbered and even-numbered paired scores and at the mean value of both scores groups, the numbers impress us of a very high consistency between the two groups of scores. In other words, when we look at the distribution of scores in both score groups we find very slim difference between the students' performance as shown in the odd-numbered group and in the even-numbered group. Therefore, we would expect to get a high reliability level of the test. However, because the scores are too far from normal distribution and are spread too largely, as indicated by the standard deviation values of 0.53 for odd-numbered items, and 0.56 for even-numbered items. The split-half reliability derived from applying pearson correlation formula is 0.42, and when converted to full-test reliability using Cronbach alpha coefficient, a reliability level of 0.60 is achieved. 0.6 coefficient point is a low reliability level given the highest point of correlation coefficient is 1. So interpretation can be made that the test developed, by moving decimal two places to the right, is only 60 per cent consistent or reliable with 40 per cent measurement error. This result of calculation make us ask ourselves a question: How could a test which seemingly has a very high internal consistency through looking at the slim difference between the two scores distributions and through the close mean value of the two score groups, but have a very low level of reliability? This unexpected result is due to the problematic use of reliability measure of NRT when applied in CRT.

Fortunately many other strategies have been worked out for investigating the consistency of criterion-referenced test—strategies that do not depend on a high standard deviation. In general these approaches fall into three categories (Berk 1984, p. 235 in Brown 2005): *threshold loss agreement, squared-error loss agreement, and domain score dependability.* These three strategies have been developed specifically for CRT consistency estimation. However, they are not described in the present article for to attempt so requires at least as many pages as have been used for this article this far which is not affordable here. For the present, suffice it to be aware of the problematic application of NRT-related reliability estimates for estimating CRT-related reliability.

Conclusion

CRT has very different use compared with NRT that Classroom language teachers should be very aware of. Its main purpose is to measure instructional objectives achievement by students. The criterion set up as reference to judge students' performance should be specified in details before translated or transformed into numbers or points as representing minimum criteria to base the students' performance judgment so that achievement decisions such, as passing or failing the test, can be convincingly made. Reliability estimate of a test developed should be afforded by applying appropriate formula. CRTs have very different characteristics from NRTs in terms of the distribution of scores and their dispersion, where CRTs do not expect the test results to approximate normal distribution of scores nor great extent of scores dispersion or a large standard deviation value. Therefore applying inappropriate procedures or formulas to measure the reliability of a CRT may tell wrong information, and thus raise a question about the quality of the test developed.

References:

Brown, Douglass. 2004. Language Assessment, Principles and Classroom Practice. San Francisco: Longman

Brown, J.D. 2005. *Testing in Language Program.* New York: McGraw-Hill

Cohen, Andrew. 1996. *Assessing Language Ability in the Classroom. Boston: Heinle & Heinle Publishers.*

Djiwandono, Soenardi. 2007. *Tes Bahasa: Pegangan Bagi Pengajar Bahasa.* Universitas Negeri Malang: Draft.

About the Author

Khaeruddin is a lecturer at Department of Medical Record and Health Information, State Health Polytechnic of Malang, Indonesia. He earned his Bachelor in Syiah Kuala University Aceh in and Master degree in State University of Malang. In 2008-2009 he joined the Visiting Scholar Program at Indiana University USA. His scholarly interest covers the areas of psycholinguistics, language testing, and translation studies. He can be contacted at fadilkhairuddin@yahoo.com.