CRT: A Better Solution of Bilingual Assessments

Hengxi Wang

School of Education, University of New South Wales, NSW, Australia
Email: hengxi.wang@unsw.edu.au
Manuscript received August 15; revised October 5; accepted November 17, 2023; published February 8, 2024

***Abstract*—This research paper argues that Criterion-Referenced Tests (CRT) are more appropriate for language assessment than Norm-Referenced Tests (NRT). While both types of standardized assessments are widely used, CRT is becoming increasingly popular in bilingual assessment due to its ability to provide meaningful information about what learners can do with the language. The paper is divided into three parts. The first part explains concepts concerning criterion-referenced testing and its critical differences with NRT. The second part elaborates on the validity, reliability, and practicality of CRTs, with examples. The final part focuses on the benefits and difficulties of implementing CRT in Chinese universities. Overall, the paper highlights the advantages of using CRT for language assessment and emphasizes its suitability for assessing second language ability.**

***Keywords*—TESOL, language assessment**

I. Introduction

In current times, Criterion-Referenced (CR) methods have been widely implemented in college-level language assessments as compared to Norm-Referenced (NR) methods. In China, two national college English tests, the College English Test (CET) for non-linguistic/English significant students and the Test for English Majors (TEM) for students majoring in English or English related subjects, are CR-based. Furthermore, colleges that base their admission process on the assessment of applicants' language proficiency set a specific criterion of language proficiency scores, measured by language proficiency tests such as International English Language Testing System (IELTS) and Test of English as a Foreign Language (TOEFL).

The work of Brindley [1] has provided insight into why the CR-method has gained widespread acceptance. Brindley argues that this method "provides meaningful information about what learners can do with the language." This essay builds upon Brindley's viewpoints by conducting a literature review and reflection on Criterion-Referenced Test (CRT). In the interest of clarity, the terms "assessment" and "test" will be used interchangeably.

The essay will be divided into three parts. The first part will compare CRT with Norm-Referenced Tests (NRT), highlighting the advantages of CRT in language assessment. In the second part, the essay will evaluate CRT based on two main assessment principles, namely validity and practicality, to demonstrate why CRT is a superior choice for language assessment than NRT. Lastly, the essay will discuss potential challenges of implementing CRT in college-level language education.

It is expected that this essay will provide insights into the reasons why the CR-method is widely preferred in college-level language assessments and why CRT is a better choice for language assessment than NRT. This essay may also serve as a valuable resource for college-level language educators, curriculum designers, and language assessment professionals in understanding the potential benefits and challenges of implementing CRT in language education.

II. Definitions of CRT and NRT

In this essay, the terms "assessment" and "test" will be considered interchangeable for the sake of clarity. The concept of criterion-referenced measurement in language testing was first introduced by Cartier in 1968 in his work "Criterion-Referenced Testing of Language Skills". Since then, it has been widely studied by scholars such as Bachman [2–4], Brown [5, 6], Davidson [7], Hughes [8], and Lynch [9], among others. However, it has only recently gained prominence worldwide with the increasing emphasis on the learners' actual ability to use the language. According to Richards [10], criterion-referenced tests measure a student's performance based on a particular agreed-upon standard of criterion. To pass the test, the student must reach this level of performance, and their scores are interpreted in relation to the criterion score, rather than the score of other students. In other words, CRT measures how well learners have mastered a particular skill and is used to determine the level of competence attained by the learner.

In contrast, the Norm-Referenced Test (NRT) is used to determine which learners are better than others [7]. The term NRT was originally coined by Glaser [11] in his work "Instructional technology and the measurement of learning outcomes." The definition of NRT has since been developed and is now defined as "the measurement approach that is concerned with determining the relative standing, or rank order, of examinees" [9]. In other words, NRT focuses on how a student's performance compares with that of other students or a norm group.

Although many scholars have different definitions of these two concepts, they are similar interpretations [12, 13]. According to Brown [5], the separation of tests by norm and criterion interpretations is becoming increasingly important in the language testing literature. The distinction between CRT and NRT is crucial in developing and analyzing various types of tests for different purposes, such as placement, diagnosis, and achievement decisions. The next section of this essay will discuss the critical differences between CRT and NRT in detail.

III. Critical Differences Between CRT and NRT

Criterion-Referenced Testing (CRT) and Norm-Referenced Testing (NRT) are the two primary forms of standardized assessments in language testing, and the distinctions between the two are often drawn in the language assessment literature. This essay argues that the primary differences between CRT and NRT lie in two areas: the interpretation of scores and test designs and constructions.

In terms of score interpretation, CRT and NRT differ in that CRT scores are interpreted absolutely, while NRT scores are interpreted relatively. The score in CRT is meaningful without reference to the scores of other examinees, while NRT scores can only be meaningful in comparison to the scores of others. Moreover, under CRT, students are judged against a set of descriptors of desired performance, while under NRT, there is no absolute criterion.

Furthermore, CRT and NRT differ in test designs and constructions. NRT tends to be much more straightforward, with test items labeled and specified in number. The multiple-choice test format is typical of NRT, which is mainly organized to test the usage of separate language skills such as vocabulary and grammar. In contrast, because criteria are central to CRT, language skill descriptions are required to be specific to clarify the criteria. The test items in CRT tend to be more open-ended and expect students' prompt responses.

Finally, it is important to note that the differences between CRT and NRT are not mutually exclusive. Both forms of assessment can be used for different purposes such as placement, diagnosis, and achievement decisions. This paper does not seek to distinguish between CRT and NRT but rather persuade readers that one type is preferable to the other.

IV. Issues of CRT

In recent years, Norm-Referenced Tests (NRT) have been criticized for their inadequacy in assessing language students' abilities, leading to increasing popularity of Criterion-Referenced Tests (CRT) in language assessments. O'Malley and Pierce [14] argue that NRT are not appropriate for language learners, and this paper aims to further explore the issues with NRT.

The first issue with NRT is the use of the "bell curve" [13] in grading, which may have negative effects on students' motivation in language learning. This often results in students forming relatively homogeneous groups [6] and perceiving themselves as "bad" students who are less able than their peers, regardless of their actual performance in the classroom. While educators may hesitate to label any student as "bad," NRT tests require teachers to compare and categorize students based on their test scores, which may not align with their actual performance in the classroom.

The second issue with NRT is its limitations in providing teachers with the information they need. Teachers need to understand their students' learning processes and progress to plan further instruction. However, the multiple-choice format of NRT assesses only receptive skills of reading and listening, ignoring the productive receptive skills of speaking and writing. This limited assessment may cause teachers to focus solely on the skills emphasized in the tests and neglect other areas of the curriculum. McNamara [13], O'Malley and Pierce [14] all express similar concerns, noting that NRT may limit the curriculum to isolated and lower-level skills.

While NRT has advantages and can identify relative strengths and weaknesses in a particular skill area, CRT is more up-to-date and better aligned with the varied needs of teaching and learning in both a narrow and broader sense. Therefore, the following section of this paper will argue in favor of CRT in terms of its validity, reliability, and practicality.

V. Strengths of CRT

In recent years, language testing has shifted its focus from solely measuring learners' scores to evaluating their actual ability to use the language. Madsen [15] has categorized the development of language testing into three stages, namely intuitive, scientific, and communicative, with the current trend emphasizing the evaluation of language use over language form. The communicative stage emphasizes learners' ability to use language in real-life situations. Given the shift in the focus of language testing, CRT has emerged as a more suitable method of assessment than traditional NRT. CRT's varied criteria, aimed at evaluating learners' performance, aligns more closely with the communicative approach to language testing. The following section aims to justify CRT in terms of its validity, reliability, and practicality, with the use of relevant examples.

A. CRT Has the Greater Validity

In the realm of language testing, the term "validity" refers to how well a test measures what it is intended to measure [8]. This paper will examine the validity of CRT in four different aspects: language ability, content validity, and consequential validity.

Firstly, CRT is considered to be a valid measure of language ability. As Bachman [2] notes, language ability can be divided into two components: language knowledge and strategic competence. CRT tests learners' strategic competence in a "real way", such as in an oral test that requires the test-taker to process and use the language effectively. This makes CRT more accurate and valid than standardized NRT, which often measures language ability with multiple-choice questions that do not fully capture the test-taker's strategic competence.

Secondly, content validity refers to the degree of correspondence between the assessment objectives and the curriculum objectives [14]. CRT objectives are expressed in the form of criteria or descriptors, which are often tied to the curriculum and teaching practices. Moreover, Lynch and Davidson [9] argue that CRT is closely related to the curriculum and teaching practices because it reflects a detailed and elaborate description of skills that should be tested. For example, when students are asked to make a writing portfolio, they are directly assessed on how much progress they have made in their writing abilities and study processes. Instructors can provide appropriate suggestions to students at different levels, and students can get a sense of achievement without being compared to other students. Therefore, CRT motivates students to learn more effectively than NRT.

Thirdly, consequential validity is the most important aspect of the test. It refers to how the assessment is used to benefit teaching and learning processes, and how it benefits students [14]. The washback effect can be an excellent example of consequential validity. According to Alderson and Wall [16], the washback effect is the extent to which the test influences language teachers and learners to do things they would not otherwise do. Hughes [8] argues that CRTs have beneficial washback effects because they can influence language learning and teaching positively. Messick [17] notes that tests with beneficial washbacks are often criterion

samples, which CRT can provide by providing clear information on how to achieve the criteria and motivating students to learn more actively. In other words, CRT is a valid and accurate measure of language ability that is closely related to curriculum and teaching practices. Furthermore, CRT can have positive washback effects on language learning and teaching, which can motivate students to learn more effectively

B. CRTs Can Be Reliable Enough

Tests are crucial to decision-making in education and other fields. The reliability of tests is of paramount importance, especially when people's lives may depend on them. For instance, international students need to take IELTS or TOEFL to study abroad. However, the underutilization of Criterion-Referenced Tests (CRT) is, to some extent, due to concerns about their reliability. Procedures for CRT, especially concerning the estimation of reliability, are not well established. Nonetheless, as Hughes [8] argues, the lack of agreed procedures for CRT is not sufficient reason to exclude them from consideration. This section suggests ways to ensure the reliability of CRT is acceptable.

Firstly, instructors can minimize the potential sources of inconsistency through test design [4]. Hughes [8] suggests some methods, such as limiting candidates' freedom. Given that CRT test items tend to be more open-ended, this aspect should be appropriately considered when developing CRT tasks. For example, if testing students' letter-writing abilities, the task can be designed as either "Writing a letter of complaint to a shop manager" or "Writing a letter of complaint to the shop manager about the broken TV set you bought yesterday. In the letter, you must cover the following points: one, describe your problem; two, explain the situation when you bought the TV set; and three, ask for a refund or repair." The second task is more reliable as it specifies the requirements in a specific range, which allows candidates to demonstrate their abilities while avoiding over-exhibition that may be difficult for examiners to evaluate.

Secondly, while objective tests such as NRT can achieve perfect reliability, the reliability of CRT is dependent on subjective judgments. Nonetheless, there are ways to achieve sufficient reliability [8]. Two ways to achieve sufficient reliability are highlighted here due to word limitations. Firstly, analytic scoring can guarantee consistency, especially for L2 students, who may vary in different language skills. Secondly, scorers should be carefully trained and regularly evaluated, as the test results depend on their subjective judgments in CRT.

Furthermore, the training of scorers can enhance the reliability of CRT. It is essential to develop a clear and detailed scoring rubric, which should be made available to all the scorers. Scorers should be trained to interpret and apply the rubric consistently. In addition, it is helpful to have a pilot test to ensure that the scoring rubric is well designed, and the scorers are competent enough to apply it effectively [8]. Although the reliability of CRT is a concern, there are several ways to ensure that it is acceptable enough for use in language assessment. Test design, scoring, training of scorers, and using multiple raters can all contribute to improving the reliability of CRT. By addressing the issues of reliability, CRT can provide a valid, reliable, and practical alternative to traditional NRT in language assessment.

C. Practicality of CRT

Define Ascertaining the practicality of Criterion-Referenced Tests (CRTs) is essential as it pertains to the feasibility of their implementation and the worthiness of their development and use [4]. Although CRTs have higher resource requirements, both in terms of facilities and human resources, their benefits far outweigh the challenges they pose.

To illustrate the practicality of CRTs, it is noteworthy to consider how educators in China previously believed that the inclusion of oral tests in the national College English Test (CET) would be impractical. The increasing population of candidates would require more time and places, and there would be a need for a comprehensive testing, selecting, and training process for the scorers, which would be time-consuming and costly. However, after careful research and design, oral tests were successfully incorporated into the CET. More importantly, this move had a massive positive washback on English teaching practices in Chinese universities. The "College English Curriculum Requirement" issued in 2004 now emphasizes more communicative language teaching methods, and oral English tasks are evident in most new textbooks and teaching materials for college English teaching.

Furthermore, another example of the practicality of CRT is the use of writing portfolios in language testing. Writing portfolios involve collecting a series of writing samples from a student throughout a course or program, which are then evaluated according to specific criteria. Although it requires more resources than NRTs, writing portfolios have several advantages over traditional tests. For example, they allow for a more comprehensive assessment of a student's writing abilities and provide more detailed feedback to the student. In addition, they encourage students to engage in the writing process more actively and reflect on their own learning progress.

In summary, while CRTs may require more resources than NRTs, their benefits outweigh the challenges they present. As demonstrated by the examples of oral tests in China and writing portfolios, CRTs have the potential to bring positive washback to language teaching and promote communicative language learning. Thus, educators should continue to explore the practicality of CRTs in language testing and consider incorporating them into their assessment practices.

VI. Benefits and Challenges in Implementing CRT in Chinese Universities

In the context of Chinese education, English language learning in secondary and high school centers primarily on the acquisition of linguistic elements. As a result, assessments in these settings tend to evaluate the mastery of language knowledge. In contrast, the goal of English language learning at the tertiary level, as outlined in the College English Curriculum of China, is to cultivate students' ability to use the language effectively. Given this shift in focus from linguistic elements to language use, Criterion-Referenced Testing (CRT) may be better suited to assess English language proficiency at the tertiary level. Nevertheless, the implementation of CRT in Chinese

universities may pose both benefits and challenges, which are discussed in the subsequent sections.

A. Benefits of CRTs in China

Criterion-Referenced Testing (CRT) may provide various benefits for English language learners and teachers in China. Firstly, it enables students to focus on developing their communicative competence in English, which is a critical skill for succeeding in today's globalized world. Since students in the Chinese education system typically study for exams, CRT can help shift their attention to the practical use of language rather than mere memorization of vocabulary and grammar rules. By testing language use instead of linguistic elements, CRT may encourage students to develop their language skills more effectively [19].

Secondly, CRT can promote a sense of achievement and motivate students to study harder. Instead of comparing students with their peers, CRT assesses their language abilities against pre-determined criteria or descriptors. Thus, students can understand what they need to achieve to demonstrate their language proficiency, which can lead to a sense of accomplishment and increase their motivation to learn [6].

Thirdly, the use of CRT in college English tests can provide valuable information for teachers to better understand their students' learning progress. Descriptors or criteria in CRT are aligned with the course objectives, enabling teachers to evaluate students' language abilities more accurately and relate test results to the curriculum more effectively. As a result, teachers can adjust their teaching methods to better cater to the students' learning needs and provide personalized feedback to students to further enhance their language skills [13]. Overall, CRT has the potential to enhance language teaching and learning in China by encouraging practical language use, promoting a sense of achievement, and providing valuable information for teachers.

B. Challenges of CRTs in China

Despite the potential benefits of Criterion-Referenced Testing (CRT) for language teaching and learning in China, there may be challenges associated with its implementation.

Firstly, CRT differs substantially from traditional Norm-Referenced Testing (NRT) in terms of test format, scoring, and aim. Resistance to CRT may come from educators who maintain the belief that the teaching of linguistic elements, such as vocabulary and grammar, represents the most "concrete" knowledge that should be tested. Some English teachers may regard teaching as limited to the analysis of reading texts, explanation of new vocabulary, grammatical points, and translation of sentences into Chinese. Consequently, these teachers may protest against tests that do not measure these "basic" language points.

Secondly, CRT requires rater training programs, which may pose a significant challenge for universities. Final exams in many universities are carried out annually using a fixed test format, which typically involves multiple-choice items, with computers performing all scoring tasks except for writing. Even writing tasks do not demand sophisticated training because the criteria for scoring are generally consistent. However, CRT requires teachers to judge students' performances against specific criteria, which can be challenging given the significant variation in language abilities among teachers. Thus, training teachers to become effective raters may prove to be a crucial challenge.

Additionally, implementing CRT in universities requires support from university authorities who have the power to decide whether to adopt the test. The implementation of CRT may necessitate significant changes in organizing teaching activities, such as revising the school syllabus, changing materials, training students and teachers, and designing new evaluation plans. This may entail a considerable expenditure of time and money. Furthermore, universities may be bureaucratic and reluctant to make changes that require significant energy and resources.

VII. Conclusion

In conclusion, it is evident that CRTs are a significant development in the field of language testing. They provide a more comprehensive approach to measuring students' language proficiency, focusing on their ability to use the language rather than just mastering its linguistic elements. The benefits of CRTs have been extensively discussed in this essay, including their ability to motivate students to learn and to provide teachers with valuable information about their students' learning process.

However, as with any assessment tool, CRTs have limitations that must be considered when implementing them. The most significant challenge in using CRTs lies in their proper implementation. If not used correctly, the negative washback can occur, and the validity of the test may be at stake. Insufficient scorer training is another challenge, which may lead to unreliable test results. Moreover, CRTs require a great deal of work to develop, implement, analyze, and revise.

Therefore, to maximize the strengths and minimize the weaknesses of CRTs, all-round considerations of the varied testing contexts are required. Developing good CRTs requires careful designing, analytic scoring, and sufficient scorer training. Educators must use CRTs correctly to avoid negative washback and maintain the validity of the test. With proper implementation, CRTs can provide a promising and valuable approach to language assessment, and the positive washback would be great. In conclusion, the appropriate use of CRTs in language assessment has the potential to significantly benefit language teaching and learning, and it is worth exploring further in the field of language testing.

Conflict of Interest

The author declares no conflict of interest.

References

[1] G. Brindley, "Defining language ability: The criteria for criteria," in *Current Development in Language Testing*, Singapore: SEAMEO Regional Language Centre, 1991.

[2] L. F. Bachman, *Fundamental Considerations in Language Teaching*, Oxford: Oxford University Press, 1990.

[3] L. F. Bachman, "What does language testing have to offer?" in *A New Decade of Language Testing Research*, 1993.

[4] L. F. Bachman and A. S. Palmer, *Language Testing in Practice*, Oxford: Oxford University Press, 1996.

[5] J. D. Brown, "A comprehensive criterion-referenced language testing project," in *A New Decade of Language Testing Research*, 1993.

[6] J. D. Brown and T. Hudson, *Criterion-Referenced Language Testing*, UK: Cambridge University Press, 2002.
[7] F. Davidson and B. K. Lynch, *Testcraft*, New Haven: Yale University Press, 2002.
[8] A. Hughes, *Testing for Language Teachers*, 2nd ed., Cambridge: Cambridge University Press, 2003.
[9] B. K. Lynch and F. Davidson, "Criterion-referenced language test development: linking curricula, teachers, and tests," *TESOL Quarterly* vol. 28, no. 4, pp. 727–743, 1996.
[10] J. C. Richards, *The Language Teaching Matrix*, New York: Cambridge University Press, 1990.
[11] R. Glaser, "Instructional technology and the measurement of learning outcomes: Some questions," *American Psychologist*, vol. 18, pp. 519–521, 1963.
[12] G. Brindley, "Defining language ability: The criteria for criteria," in *Current Development in Language Testing*, Singapore: SEAMEO Regional Language Centre, 1991.
[13] T. McNamara, *Language Testing*, Oxford: Oxford University Press, 2000.
[14] J. M. O'Malley and V. V. Pierce, *Authentic Assessment for English Language Learners: Practical Approaches for Teachers*, U.S.A.: Addison-Wesley Publishing Company, 1996.
[15] S. H. Madsen, *Techniques in Testing*, Oxford: Oxford University Press, 1983.
[16] J. C. Alderson and D. Wall, "Does washback exist?" *Applied Linguistics*, vol. 14, pp. 115–129, 1993.
[17] S. Messic, "Validity and washback," in *Language Testing*, 1996.
[18] College English Curriculum (Trial). (2004). [Online]. Available: http://211.151.90.18/info/107/255.asp
[19] K. Morrow, "Evaluating communicative tests," in *Current Development in Language Testing*, Singapore: SEAMEO Regional Language Centre, 1991.