Criterion-Referenced Language Testing

Research and Development Center for Higher Education Paul Westrick

The normal distribution, also known as the bell curve, is a feature of the natural world around us. If we measure the weights of all the twenty-year old males at the in the country, we can expect them to fall in a normal distribution. If we measure the heights of all the twenty-year old females in the country, we can expect them to fall in a normal distribution. We can measure a variety of things, and if we measure an entire population, or at least a large sample of the entire population, we can expect to see a normal distribution. With norm-referenced tests (NRTs), such as the TOEFL and TOEIC tests, a normal distribution is the desired outcome of the test scores, but when it comes to criterion-referenced tests (CRTs), also referred to as achievement tests and classroom tests, the normal distribution is not desired, for it is hoped that during a course students learn the material specific to the course and complete assigned tasks. Over the years norm-referenced testing principles have been misapplied to criterion-referenced testing situations, and what follow are a brief history of this problem and an overview of how norm-referenced testing and criterion-referenced testing differ in regard to their purposes, contents, structures, and score distributions and interpretations.

With the rise of standardized testing in the early 20th century, some educators became enamored with features of norm-referenced testing and forgot the original purpose of classroom testing — determining if students had learned the course material. While some were entranced by bell curves and reliability estimates, others were not, and over the years criticisms of the use of norm-referenced testing principles in classroom assessments grew in number. As Hopkins, Stanley and Hopkins (1990) describe it,

> The CRT movement is in part a reaction to the misuse of psychometric methods (especially reliability theory) that were developed for assessing individual differences in aptitudes and abilities.... NRT oriented persons were often inordinately interested in achievement tests with high reliability coefficients. Items were eliminated or retained for future use solely on the basis of their discrimination or difficulty. Since items that are answered correctly (or incorrectly) by a large percentage of the examinees tend to have lower discrimination indexes, they were often eliminated solely on that empirical criterion.... In other words, many achievement test developers lost sight of content validity — that the *items must first and foremost be representative of the domain (content) and possess objectives to be assessed, and focused on high reliability as an end in itself.* (pp.184-185 emphasis in original)

By focusing on normal distributions and reliability coefficients, classroom testing had become warped. It had come to be that, in the minds of some people, if too many students in a given class studied hard, learned everything taught in the course, and received high test scores, there was something wrong because the test scores did not fall in a normal distribution. Likewise, when too many diligent students completed all their assignments as specified by their teacher, it caused problems because the grades did not fall in a normal distribution. At the other extreme, in classes filled with laggards, students with very poor comprehension of the course material could still get high grades because most of their peers received even lower marks on the tests

and assignments. In either case, whether a student was studious or lazy, when teachers graded on a curve, the normal distribution, the grade a student received often depended more on the performance of his or her classmates than it did on his or her comprehension of the material or completion of assignments.

Fortunately, there were others who realized that the obsession with normal distributions and reliability coefficients had gone too far and in fact had no place in classroom testing. Writings on criterion-referenced testing first appeared in the early 1960s, and the topic made its way into language testing literature toward the end of that decade (Brown, 1996). Over the past four decades, the idea that there are two main categories of tests, NRTs and CRTs, has become widely accepted. Below is a brief description of the differences between NRTs and CRTs — in regard to their purposes, contents, structures, and score distributions and interpretations — drawn largely from Brown (1995, 1996) and Brown and Hudson (2002) with some additional points on drawn from other testing experts.

Purpose

The purpose of an NRT is to spread students out as much as possible so that differences, often in ability, between students can be seen. These tests are predominantly used for admission and placement purposes, situations in which it is desirable to put students in groups, and the distinctions between these groups need to be as clear as possible. They are designed to measure aptitude or proficiency — encouraging learning is not the designers' basic intent.

Conversely, the purpose of a CRT is to encourage and measure learning. Clearly defined objectives and standards are presented to the students at the beginning of the course, and tests are given to determine how much course material a student has learned and/or tasks are assigned and evaluated to see if they have been completed to established standards. When students know that their final grade will not be based on a bell curve, and when they know that if they meet the standards set at the beginning of the course they can receive the highest possible grade, they realize that each person controls his or her own destiny in the class. As Hughes (2003, p.21) writes, "Criterion-referenced tests therefore have two positive virtues: they set meaningful standards in terms of what people can do, which do not change with different groups of candidates, and they motivate students to attain those standards." Furthermore, students are competing with themselves, not their peers (McNamara, 2000). In a course in which the students admitted can reasonably be expected to be able to master the course material and/or complete the tasks, and grades are based on mastery of content and/or the satisfactory completion of tasks, not based on a curve, students can only blame themselves for poor marks.

Content

The content of an NRT should be very general, and students should have no idea of what will be on the test. As students are expected to come from various backgrounds (different classes, schools, communities, even countries), there should be a wide variety of test items.

With a CRT, the content should be drawn from the course, and students should have a good idea of what

they will encounter on the test. This does not mean that they should know the questions in advance, but there should be no surprises either. The test content should come from what was covered in class and/or in the outside reading (or listening, or viewing) assignments. By testing only what was covered in the course, particularly what was emphasized by the instructor, and by not including extraneous material for the sake of lowering scores and/or creating a curve, it reinforces in the students' minds the importance of the material from the classes and the outside assignments. Similarly, the evaluation of assigned tasks should be based on established standards, standards that should not change even if all students (or no students) are meeting them. Testing experts and students generally find these practices to be fair and objective.

Structure

NRTs should be long, and they may include lengthy subtests. The structure of an NRT is closely linked to its purpose, which is to distribute scores over a broad range. To accomplish this objective, it is essential that there are a large number of items on the test. These tests typically take hours to administer because they are so large.

CRTs tend to be shorter than NRTs. This is largely due to CRTs being narrowly focused on the content of particular courses. CRTs may have various subtests, and these, too, will be short in length and focus on specific material from the course. As they are usually administered during class hours or during the examination period at the end of a term, they tend to be much shorter in duration than NRTs.

Score Distributions and Interpretations

These interrelated issues, score distributions and interpretations, are where NRT principles are often mistakenly applied to CRT situations. With NRTs, there should be a normal distribution of scores, a bell curve, and a student's performance is compared to those of other students. A student's performance is relative to the performances of others, and scores are typically reported in percentiles ― students are ranked. With CRTs, distributions may be normal, but they are usually not normal. A student's performance is judged regarding the course material covered or tasks assigned. A student's scores are absolute in that they pertain to how much the student has mastered or completed, and they are generally reported as a percentage ― there is no ranking, and every student can potentially get a perfect score.

These distinctions between the score distributions and interpretations of NRTs and CRTs have been thoroughly covered in the educational and language testing literature. As explained by Hopkins et al. (1990),

> Individual differences are the major emphasis of norm-referenced testing (NRT), but they are of no concern in mastery or criterion-referenced testing (CRT). If everyone scores 100 percent on the test, so much the better (assuming the test is valid). CRT assessments should reveal what competencies an individual student does and does not possess, not how he or she compares with norms or peers (NRT). (p.184)

In Bachman's (1990) words,

> Criterion-referenced (CR) tests are designed to enable the test user to interpret a test score with reference to a criterion level of ability or domain content. An example would be the case in which students are evaluated in terms of their relative degree of mastery of course content, rather than with respect to their relative

> ranking in the class. Thus, all students who master the course content might receive an 'A', irrespective of how many students achieve this grade. The primary concerns in developing a CR test are that it adequately represent the criterion ability level or sample the content domain, and that it be sensitive to levels of ability or degrees of mastery of the different components of that domain. (p.74)

Brown (1995) writes,

> Teachers will be comforted to know that a normal distribution (commonly known as a bell curve) may not necessarily occur in the scores of their classroom tests.... In addition, on CRTs, the ideal distributions would occur if all of the students scored zero at the beginning of a course (indicating that they all desperately needed to learn the material) and 100 percent at the end of the course (indicating that all of the students have learned all of the material perfectly). Neither of these ideals is ever really met, even with a good test, but the scores might logically be "scrunched up" toward the bottom of the range at the beginning of a course and toward the top of the range at the end of the course. Hence for a number of reasons, expecting a normal distribution in classroom testing is unreasonable. Nonetheless some administrators expect just that, usually in the name of "grading on a curve." (p.17)

As the practice of task-based teaching and assessment has grown in the field of language teaching, the thoughts of Hughes (2003) on CRTs should be noted as well:

> The purpose of criterion-referenced tests is to classify people according to whether or not they are able to perform some task or sets of tasks satisfactorily. The tasks are set, and the performances are evaluated. It does not matter in principle whether all the candidates are successful, or none of the candidates successful. The tasks are set, and those who perform them satisfactorily 'pass'; those who don't, 'fail'. This means that students are encouraged to measure their progress in relation to meaningful criteria, without feeling that, because they are less able than most of their fellows, they are destined to fail. (p.21)

In conclusion, NRTs and CRTs are different instruments. NRTs are for comparing students. They are typically lengthy tests with numerous items that are general in nature. Students' scores should be spread out over a continuum, and the distribution should be normal. Students' scores are usually reported as a percentile as the purpose of the test is to compare students' performances. CRTs are for determining how much course material has been mastered or how many tasks have been completed. They are generally shorter in length than NRTs, and the items or tasks are specific to the course. Student scores may not fall in a normal distribution, particularly if a class consists of diligent students who master the course materials and/or complete all tasks to established standards. Scores are usually reported as percentages or as pass/fail as the purpose of the assessment is to check mastery of content or completion of tasks. Educators need both NRTs and CRTs, but it is important to understand that the principles of norm-referenced testing only apply to norm-referenced testing situations and that the principles of criterion-referenced testing only apply to criterion-referenced testing situations. They are not interchangeable.

References

Bachman, L. (1990). *Fundamental considerations in language testing.* Oxford: Oxford University Press.

Brown, J. D. (1995). Differences between norm-referenced and criterion-referenced tests. In J. D. Brown & S. O. Yamashita (Eds.), *Language testing in Japan* (pp.12-19). Tokyo: Japan Association for Language Teaching.

Brown, J. D. (1996). *Testing in language programs.* Upper Saddle River, NJ: Prentice Hall Regents.

Brown, J. D. & Hudson, T. (2002). *Criterion-referenced language testing.* Cambridge: Cambridge University Press.

Hopkins, K., Stanley, J. & Hopkins, B. R. (1990). *Educational and psychological measurement and evaluation* (7th ed.). Englewood Cliffs, NJ: Prentice Hall.

Hughes, H. (2003). *Testing for language teachers* (2nd ed.). Cambridge: Cambridge University Press.

McNamara, T. (2000). *Language testing.* Oxford: Oxford University Press.