Questions and answers about language testing statistics:

Differences in how norm-referenced and criterion-referenced tests are developed and validated?

James Dean Brown
brownj@hawaii.edu
University of Hawaiʻi at Mānoa

Question:

What are the major differences between norm-referenced and criterion-referenced tests? How can these two tests be best developed and validated? [Submitted by a participant in the Kuroshio (Aloha Friday) Seminar that Kimi Kondo-Brown and I conducted on May 23, 2014 at the Bunkyo Civic Center in Tokyo]

Answer:

I have discussed the major differences between norm-referenced and criterion-referenced tests in a number of places (most recently in Brown, 2012a). So I will only touch on those differences briefly here. I have also explained at length the different strategies that should be applied in developing and validating the two families of tests in a number of places. However, I have never summarized those different strategies side-by-side in one short and straightforward article. I will attempt to do just that here by addressing the following sub-questions: What are the differences between the norm-referenced and criterion-referenced families of tests? What strategies are used to develop and validate NRTs and CRTs? What are the differences in NRT and CRT development and validation strategies?

What are the Differences Between the Norm-Referenced and Criterion-Referenced Families of Tests?

Norm-referenced tests (NRTs, sometimes referred to as *standardized tests*) and criterion-referenced tests[1] (CRTs, also known as *classroom tests*) are two families of tests that are distinguished most clearly in terms of the ways scores are interpreted, the purposes of the tests, levels of specificity, the distributions of scores, the structures of the tests, and what we want the students to know in advance. In more detail, the two types of tests differ in:

- *The ways scores are interpreted* differ is that NRTs are designed to compare the performances of students to one another in relative terms, while CRTs are built to identify the amount or percent of the material each examinee knows or can do in absolute terms.
- *The purposes of the tests* also differ with NRTs primarily designed to spread examinees out on a continuum of general abilities so examinees' performances can be compared to each other

[1] Note that, since the question addressed to this column was clearly written by a person interested in testing, but primarily a teacher, the types of CRTs I am referring to here are not the formal subcategory of CRTs known as domain-referenced tests (which tend to be large scale), but rather those CRTs used by teachers on a more focused classroom level.

(usually with standardized scores), while CRTs are designed to assess the amount of material that the examinees know or can do (usually expressed in percentages).

- *Levels of specificity* are necessarily different with NRTs tending to measure very general language abilities (for proficiency or placement purposes), while CRTs usually focus on specific, well-defined (and usually objectives-based) language knowledges or skills (for diagnostic or achievement purposes).
- *The distributions of scores* also differ in that, ideally, NRT scores are normally distributed (indeed items are selected to ensure this is the case), while CRT scores ideally would produce quite different distributions at different times in the learning process: with students scoring very low in a positively skewed distribution at the beginning of a course on a diagnostic CRT (indicating that they needed to learn the material) and students scoring generally high in a negatively skewed distribution at the end of the course on an achievement CRT (indicating that most of them mastered the material; indeed, in the unlikely event that all students master all the material, they should all score 100%).
- *The structures of the tests* also differ with NRTs tending to have many items with a few long subtests (e.g., listening, grammar, reading, etc.) each of which has diverse item content, while CRTs are typically built around numerous, short subtests that contain well-defined and similar items in each.
- *What we want the students to know in advance* of the test differs in that, for NRTs, security is usually an important issue because we do *not* want examinees to know the content of the test items, while for CRTs, we teach the content of the course and want the students to study that content, so we tell them what to study, and we test that content. If they know the content, they should succeed.

What Strategies Are Used to Develop and Validate NRTs and CRTs?

Table 1 summarizes the strategies used to develop NRTs and CRTs in two separate columns. I hope that this table is clear without any direct explanation. Nonetheless, some discussion of the differences between NRT and CRT development strategies will be provided below.

Table 1
Strategies Used to Develop NRTs and CRTs

Steps	***NRT (Standardized)***	***CRT (Classroom)***
1. Plan test	Plan based on test specification/blueprint and general item specifications.	Plan with course objectives developed and in hand; when possible, using item specifications will help.
2. Create items	Create a large pool of items at about the right level of difficulty in the general area being tested (e.g., reading comprehension).	Create about 10 items that measure what the students should be able to do on each of the course objectives (say objectives 1-9) at the end of the course; divide the items into two forms of the test, say forms A and B such that there are about 5 items on each test for each of the 9 objectives/subtests.
3. Edit items	Use item writing guidelines like those found in Brown (2005, Chapter 3) to carefully proofread and improve all items.	Use item writing guidelines like those found in Brown & Hudson (2002, Chapter 3) to proofread and improve all items. Perform item congruence and applicability analysis (as described in Brown & Hudson, 2002, pp. 98-100) to make sure items match objectives.

4. Pilot items	Pilot the items *with a single large group of examinees* that has the same characteristics and range of abilities as the examinees in the ultimate test group (e.g., if the test is being developed for proficiency purposes, pilot it with a large group of students ranging from near-zero English to near-native; if the test is for placement purposes at a specific institution, the test should be piloted with examinees in the narrower range of abilities found there).	Ideally, pilot the two forms *at the beginning of the course* as diagnostic tests (with half of the students randomly selected to take each form); score and give the students diagnostic feedback objective-by-objective based on the subtests. Then, administer the same tests *at the end of the course* as achievement tests such that students who took Form A at the beginning take Form B at the end, and vice versa; include the scores in the students' grades, but keep the tests for further analysis.
5. Analyze items	Calculate *item facility* (*IF* = the proportion of examinees who answered each item correctly) and *item discrimination* indexes (*ID* = proportion of examinees in the upper third on the whole test who answered each item correctly minus the proportion in the lower third) (see Brown, 2005, pp. 66-76).	Calculate *difference indexes* (*DI* = proportion of students who answered each item correctly at the end of the course minus the proportion at the beginning) and *B indexes* (*BI* = proportion of those examinees who passed the whole test that answered each item correctly minus the proportion of correct answers for those students who failed) (see Brown, 2005, pp. 76-84, or Brown & Hudson, 2002, pp. 118-148).
6. Select items	Revise the test by selecting those items with the highest *ID* values while keeping an eye on the *IF* values to adjust the difficulty of the test up or down as necessary.	Revise the test by selecting those items with the highest *DI* values within each objective/subtest (perhaps the best 3 out of 5). If *DI* values are not available, select the highest *BI* values in each objective/subtest (again, perhaps the best 3 out of 5).
7. Revise test	Create a new, shorter, more efficient revised test based on the item analyses and selection in Steps 5 and 6 for future proficiency or placement purposes.	Create new, shorter, more efficient, revised Forms A and B based on the item analyses and selection in Steps 5 and 6 for future use as diagnostic and achievement tests.

Table 2 summarizes the strategies used to validate NRTs and CRTs in two separate columns. Again, this table should stand alone as a summary, but further discussion will be provided in the next section.

Table 2
Strategies Used to Validate NRTs and CRTs

Steps	***NRT (Standardized)***	***CRT (Classroom)***
8. Examine consistency	Study the *reliability* of scores by using test-retest, parallel forms, or internal consistency strategies–the most commonly applied internal consistency estimates are Cronbach alpha, K-R20 or K-R21 (for full explanations of all these reliability strategies, see Bachman, 2004, pp. 153-191; Brown, 2005, pp. 169-198; Brown, 2013a).	Study the *dependability* of scores by using threshold loss agreement (agreement or kappa), squared error loss (Φ_{λ}), or domain score dependability (Φ) strategies. If resources are limited as in most classroom settings, teachers can use the K-R21 reliability statistic as a conservative estimate of Φ mentioned above (for full explanations of these dependability strategies, see Bachman, 2004; pp. 192-205; Brown, 2005, pp. 199-219; Brown, 2013b).
9. Examine validity	Use evidential strategies, which include the traditional content, construct, and criterion-related validity strategies. Also use the more recently developed consequential strategies including examination of the values implications and social consequences of score interpretations and uses (see Bachman, 2004, pp. 257-293; Brown, 2005, pp. 220-248).	Use the only evidential strategy that typically makes sense for CRTs, which is the traditional content validity approach. Teachers may also want to use the more recently developed consequential strategies that take into the account values implications that they are expressing by the choices they make in test design as well as the social consequences of their score interpretations and uses (see Brown, 2012b; Brown & Hudson, 2002, pp. 212-268).

What are the Differences in NRT and CRT Development and Validation Strategies?

Careful examination of Table 1 will reveal key differences between NRT and CRT development strategies. In Step 1, the primary difference in *test planning* is that CRTs are more specific and objectives-based, while NRTs are more general. In Step 2, the difference in *creating items* is that a more general pool of items is developed for NRTs, but smaller, more specific item pools are created for each objective/subtest in CRTs. In Step 3, *editing items* includes using item guidelines for both types of tests, but item congruence and applicability analyses are key to CRT development. In Step 4, the key difference in *piloting items* is that NRTs can be piloted in one shot and must include the whole range of abilities being tested, while CRTs are best piloted at the beginning and end of appropriate instruction and should focus only on what is being taught. In Step 5, the key difference is that *analyzing items* for NRTs is based on *ID*, and *IF* (in that order), while ideally, CRT item analysis is based on *DI*, but in a pinch can be based on *BI*. In Step 6, the key difference in *selecting items* is that, for NRTs, it is based on the highest *ID*s, and then on *IF* (to adjust test difficulty), while ideally CRT item selection is based on the highest *DI*s, but in a pinch on the highest *BI* values. In Step 7, the prime difference in *test revising* is that the ultimate product for NRTs is typically one large general test (or sometimes large subtests like grammar, listening, reading, etc.), but for CRTs, the resulting product is usually a collection of small, focused, objectives-based subtests, ideally in two forms

Table 2 reveals key differences between NRT and CRT validation strategies. In Step 8, the NRT *reliability* practices listed in the table are those laid out and explained for NRTs in most language testing (or more general testing) books. For CRTs, the *dependability* procedures shown in the table can clearly become quite elaborate. However, teachers need only address the common sense questions of whether the scores on their tests are consistent, fair, and consistently represent the knowledge and abilities of all students. If resources are limited as is the case in most classroom settings, teachers can use the K-R21 reliability estimate as a conservative estimate of domain-score dependability (Φ) referred to in the table (see argument for this strategy in Brown, 2005, p. 209).

In Step 9, the *validity practices* for NRTs listed here are also those laid out and explained for NRTs in most language testing (or more general testing) books including evidential strategies like content, construct, and criterion-related validity strategies and consequential strategies examining values implications and social consequences. For CRTs, the content validity approach listed in the table is the only one that always makes sense; it involves systematically analyzing and assessing the degree to which test items are measuring what the teacher is claiming to test, often by laying out the test items side-by-side with the course objectives (and with the teaching materials nearby for reference) and systematically comparing items to objectives. There are three key questions that teachers may want to consider in this regard (note that these questions and those in the next paragraph are adapted from and explained more fully in Brown, 2012b, 2013c):

1. Does the content of my test match the objectives of the class and the material covered?
2. Do my course objectives meet the needs of the students?
3. Do my tests show that my students are learning something in my course?

Teachers may also want to consider the *values implications* of their testing, scoring, and decision making by addressing some or all of the following questions: How do the learning/teaching values that underlie my test design, the resulting scores, and the decisions based on them match my beliefs and values? The beliefs and values of my students? Their parents? My colleagues? My boss? Etc.? Teachers may also want to think about the *social consequences* of their scores and decisions by addressing the following questions: What will happen to my students as a consequence of the decisions I make based on these test scores? Is this a small-stakes decision that is only a small part of a course grade, or will this test have larger consequences for students (e.g., determine whether or not the student passes the course, graduates with a diploma, etc.)?

Conclusion

In answering the question posed at the top of this column, length restrictions limited me to summarizing the differences in characteristics, development steps, and validation strategies used for NRTs and CRTs. I hope that this overview will nonetheless prove useful to readers and that anyone who wants more in-depth coverage of any aspect of these differences will be able to use the citations and references provided here to continue exploring these and related topics. I especially hope that this explanation will help practicing language teachers realize that most of the testing they do in the classroom ought to be CRT and that this column along with Brown, 2013c (which discusses solutions to problems that teachers often have with their classroom testing) will help them do a better job of assessing their students.

References

Bachman, L. F. (2004). *Statistical analyses for language assessment.* Cambridge: Cambridge University.

Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment* (New edition). New York: McGraw-Hill.

Brown J. D. (2012a). Choosing the right type of assessment. In C. Coombe, S. J. Stoynoff, P. Davidson, & B. O'Sullivan (Eds.). *The Cambridge guide to second language assessment* (pp. 133-139). Cambridge: Cambridge University.

Brown, J. D. (2012b). What teachers need to know about test analysis. In C. Coombe, S. J. Stoynoff, P. Davidson, & B. O'Sullivan (Eds.), *The Cambridge guide to language assessment* (pp. 105-112). Cambridge, Cambridge University.

Brown, J. D. (2013a). Classical theory reliability. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 1165-1181). Oxford, UK: Wiley-Blackwell.

Brown, J. D. (2013b). Score dependability and decision consistency. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 1182-1206). Oxford, UK: Wiley-Blackwell.

Brown, J. D. (2013c). Statistics Corner. Questions and answers about language testing statistics: Solutions to problems teachers have with classroom testing. *Shiken Research Bulletin*, *17*(2), 27-33. Also retrieved from the World Wide Web at http://teval.jalt.org/sites/teval.jalt.org/files/SRB-17-2-Brown-StatCorner.pdf

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

Where to Submit Questions:

Please submit questions for this column to the following e-mail or snail-mail addresses: *brownj@hawaii.edu*. Your question can remain anonymous if you so desire.

JD Brown
Department of Second Language Studies
University of Hawai‘i at Mānoa
1890 East-West Road
Honolulu, HI 96822
USA