Statistics Corner:

Solutions to problems teachers have with classroom testing

James Dean Brown
University of Hawai‘i at Mānoa

Question:

I sometimes feel like I must be making lots of mistakes when I write tests for my students. What worries me most is that I may be wasting my time and theirs because I don't know what I am doing. Can you help me by explaining common mistakes that teachers make when they design tests and how to avoid them?

Answer:

The problems that test designers have when writing and developing standardized tests (norm-referenced tests) are discussed in many language testing books. However, the problems that teachers have in implementing classroom tests (criterion-referenced tests) are rarely covered. Yet surely, testing occurs more often in language classrooms than in standardized language testing settings. So I will be happy to address the classroom testing problems that teachers face and offer solutions to those problems—at least to the best of my ability. I will do so in three sections about problems that teacher may have in test writing practices, test development practices, and test validation practices.

Test Writing Practices

In test writing practices, teachers sometimes have problems with: creating good quality test items, organizing those items in the test, and providing clear headings and directions.

Create good quality test items. The biggest single *problem* that most teachers have with tests is the tendency to treat tests as an afterthought, waiting until the last possible moment to write a test for the next day. This habit leaves teachers with too little time to create good quality items. In many cases, I suspect that this tendency is caused by lack of training in item writing, training that, if nothing else, would teach them that writing good test items takes time.

Clearly, the *solution* to this problem is to get ahold of a good book on language testing and read up on what good quality items are and how to write them (see e.g., Brown, 2005, pp. 41-65; Brown & Hudson, 2002, pp. 56-100; or Carr, 2011, pp. 25-45, 63-101). Then when it is time to write a test, make sure to allot enough time for writing good quality test items by starting early. These strategies will pay off handsomely because a carefully written test will always be better than a shoddily written one.

Organize the items. The *problem* here is that tests sometimes seem like a disorganized hodge-podge. Any test will be clearer to the students and easier for them to negotiate if the items are clearly organized into sections that make up the whole test. Teachers naturally try to organize their tests, but this can always be done better.

The *solution* is to follow at least three basic principles: (a) group items that are testing the same language point together, (b) collect items of the same format (e.g., multiple-choice, true-false, matching, writing tasks, etc.) together, and (c) group items based on reading or listening passages together with the passages they are based on. Unfortunately, these three principles are sometimes in conflict. For instance, it would be reasonable to have a test with say three reading passages; each reading passage might have one multiple-choice main-idea item, one fact item, one vocabulary item, and one inference item, and each passage might be followed by an open-ended critical-thinking item that students must answer in writing. Clearly, such a test would be following principle (c) above but not (a) and (b). Another section on the same test might group multiple-choice questions together with five for articles, five for prepositions, five for copula, and so forth. That would be following principles (a) and (b) but not (c). I stand by these three principles, but they are not hard and fast rules, and they may not all apply at the same time. Common sense should guide which of the three need to be applied and in what combinations.

Provide clear headings and directions. The *problem* is that even when a test is well-organized, the students may not understand that organization, or worse yet, they may not realize exactly what they have to do on the test. Any test will be clearer to the students and easier for them to negotiate if it has clear headings and directions.

The *solutions* involve making sure the headings are distinct from the rest of the text (in the sense that they are italicized, made bold, or otherwise emphasized) and ensuring that they clearly indicate heading levels with different forms of placement and emphasis like those used in this article (left-justified title-caps and bold italics used for main headings and beginning of paragraph first letter cap with a period and bold italics for second-level headings).

In addition, given that the students taking these tests are usually second language speakers of the target language, the directions should probably be in the students' mother tongue, or if that is not possible, the directions should be simple and direct in the target language (with clear options for asking the teacher for further clarification). Two types of directions will often serve best: general and specific directions. General directions typically provide information to students about the overall test and apply to all sections of the test. Specific directions are particular to the section for which they are supplied. One thing to keep in mind: if the phrase or sentence appears in all of the specific directions, it probably belongs in the general directions.

Test Development Practices

In test development practices, teachers sometimes have problems with: proofreading the test, using a sufficient number of items, and examining student performances on the items.

Proofread the test. Another *problem* is that, even when a good deal of effort has gone into writing good quality items, clearly organizing those items, and providing clear headings and directions, other problems may still persist including typos, spelling errors, unclear formatting, and other problems that will make the test harder for the students to understand.

The *solution*, or at least a partial solution, is to carefully proofread the test several times even though you think you have finished it. I like to proofread my way through the test in different ways: reading from left to right on each line, then right to left; reading from top to bottom, and then bottom to top; I even throw the paper on the floor and look it over while standing above it (especially for logical formatting, e.g., making sure each item is on one page, that each reading passage is visible at the same time as the items associated with it; etc.). The trick is to look at the test from various perspectives because that will help in spotting typos and other problems before the tests are reproduced and handed out to students.

I also find that it helps to get others involved in the proofreading process because of the different and useful perspectives they may bring to the task. What I am suggesting is that you have a colleague, a former student, or even a spouse also proofread the test. You will be amazed at the sorts of problems they will uncover because their different perspectives on the test allow them to see things you are too close to the test to notice. Remember that, ultimately, when you administer the test to say 20 students, you will also have 20 people proofreading your test—people who are more than willing to point out a mistake that the teacher made in writing the test.

Use a sufficient number of items. The *problem* that some teachers create is that they try to test their course objectives with too few items. It stands to reason that more observations of a given phenomenon will be more accurate than fewer observations. This principle is well established in the sciences. However, even in language testing, common sense tells us that testing students with one multiple-choice item would not be reliable or accurate, indeed it simply wouldn't seem fair. Would two items be better? Or 3 items, or 10? So the principle that more items are generally better makes sense. The only real question is how many items are necessary to make the assessment of students reliable, accurate, and fair. The answer to that question depends on how good the items are. If the items are of good quality and suitable for the students in terms of their general proficiency level and what they are being taught, then fewer items will be necessary.

One *solution* is to make sure you have enough items to start with (say 50% more than you think you will need) so you can get rid of some items if they don't work very well. How many items should you have on your test? That will depend on common sense and thinking about the time constraints and the types of things you are asking your students to do on the items. So the end number will be different for each situation. But this I know, more items will generally do a better job of measuring what your students can do, but you can get away with fewer items if they are good ones.

Examine students' performances on the items. Another *problem* that arises for teachers is that they do not analyze their students' performance on their test items, much less revise those items. As a result, such teachers continue to use the same items or types of items over and over again even though those items do not work very well. You have probably found yourself in situations administering a test, when suddenly a student asks if there are two possible answers for number 11, and you realize she is right; then another student asks if any answer is correct for number 25, and you realize that there really isn't. So you tell the students to select the "best" answer, which essentially means that you recognize that there are problems with those items, and perhaps others. The next semester you are using the same test, when suddenly a student asks if there were two possible answers for number 11, and you instantly realize that you forgot to fix the items that had problems, even though students had helped you spot those problems.

One obvious *solution* is to carefully listen to students questions and comments about your test and take notes, then, after scoring the test, immediately take a few minutes to revise the test and save that version in such a way that you will remember to use it the next time you test the same material.

A more *systematic solution* would be to consider the first administration of any test a pilot run. You can then analyze the results statistically and revise on the basis of what you learn from the analysis. The actual item analyses that are probably most appropriate for classroom tests are called the *difference index*, which "shows the gain, or difference in performance, on each item between the pretest and posttest" (Brown, 2003, p. 18) and the *B index*, which "shows how well each item is contributing to the pass/fail decisions that are often made with CRTs" (p. 20). These item analysis statistics are both based on the simple percentage of students who answered each item correctly at different times or in different groups. Using these statistics and common sense, you can select those items that are most closely related to what your students are leaning in your course, replace any items that are not closely related, and make fairer decisions based on your test scores. For more about the steps in calculating and interpreting these classroom-test item statistics, see Brown (2003, 2005). If you take the time to do item analysis every time you administer a test, your tests will continue to get better every time you use them.

Test Validation Practices

In test validation practices, teachers sometimes have problems with: reporting the scores as percentages, checking the reliability of the test, and thinking about the validity of the test.

Report the scores as percentages. The first validity-related *problem* is that some teachers report the number of items answered correctly to students along with information about the distribution of scores (e.g., the high and low scores, the number of students at each score, etc.). Teachers probably do this because they (and their students) are thinking in terms of the bell curve. This approach will lead students to think competitively in terms of how they did relative to other students, rather than to how much learning they were able to demonstrate on the test.

The *solution* is a simple one. In order to encourage the students to think about how much they have learned, report their scores as percentages and explain to them that the scores reveal what proportion they learned of the material taught in the course. Your score report will be even more informative if you can give students their percentage scores for each section of the test or for each objective in the course. The important thing to keep in mind for yourself and your students is that your classroom tests are designed to measure their learning in the course (criterion-referenced testing), not to spread them out on a continuum (which is norm-referenced testing like that done on standardized tests).

Check the reliability of the test scores. The *problem* here is that some teachers fail to think about or check the degree to which they might be making decisions about their students (grading, passing/failing, etc.) based on unreliable information. What does reliability mean when it comes to test scores? Reliability can be defined as the degree to which a set of scores are consistent. This concept is important because teachers generally want to be fair and make decisions for all students in the same way. If the scores on a test are not consistent across time, across items, or especially across students, then the decision making may not be the same each time for all students. Thus reliability is really a question of fairness.

One *solution* to this reliability issue is for teachers to think about reliability in terms of *sufficiency of information*: “What teachers really need to know, from a reliability perspective, is, ‘Do I have enough information here to make a reasonable decision about this student with regard to this domain of information?’ The essential reliability issue is: Is there enough information here?” (Smith, 2003, p. 30). While Smith was pondering the idea of creating a reliability index for such an interpretation, teachers might simply ask themselves one question: do I have enough good quality information from these test items to make responsible decisions about my students?

Another *solution* to this reliability issue would be to directly address the question: To what degree are the scores on my test reliable? This could be addressed by calculating a reliability coefficient. These coefficients typically range from .00 to 1.00, which can be interpreted as a range from zero reliability to 100% reliability. Thus if a coefficient for a set of scores turns out to be .80, that means that the scores are 80% reliable (and by extension 20% unreliable). So generally, the higher this value is the more reliable the scores are. Most reliability estimates were designed for standardized tests and are not appropriate for classroom testing, but one such estimate, the Kuder-Richardson formula 21 (known affectionately as K-R21) is appropriate for classroom testing (as explained in Brown, 2005, p. 209). Calculating this coefficient is relatively easy, requiring only that the teacher first calculate the mean (*M*), standard deviation (*SD*), and number of items (*k*) (all of which can be calculated fairly easily in the Excel spreadsheet program), then enter these values into Walker's calculator for K-R21 at:

http://www.cedu.niu.edu/~walker/calculators/kr.asp

The result will be a reliability coefficient that the teacher can interpret as an indication of the consistency of the test scores involved.

Think about common sense validity issues. Another *problem* that some teachers have is that they fail to consider the validity of their test scores. Validity has traditionally been defined as the degree to which a set of test scores is measuring what it was intended to measure. In recent years, language testers have expanded their thinking about validity to include issues related to the consequences and values implications of how those scores are used.

Classroom teachers who wish to address validity issues need not get involved in learning elaborate theories or statistical procedures. They can instead start by asking themselves the following relatively simple questions (adapted from and explained more fully in Brown, 2012):

1. How much does the content of my test items match the objectives of the class & the material covered?
2. To what degree do my course objectives meet the needs of the students?
3. To what degree do my test scores show that my students are learning something in my course?
4. Will my students think my test items match the material I am teaching them?
5. How do the values that underlie my test scores match my values? My students' values? Their parents' values? My boss' values? Etc.?
6. What are the consequences of the decisions I base on my test scores for my students, their parents, me, my boss, etc.?

Your answers to the above questions will probably be matters of degree, but they will nonetheless help you understand the degree to which your test scores are valid.

Conclusion

In this column, I have explored some of the problems that teachers may face in their classroom testing in terms of test writing practices, test development practices, and test validation practices. These notions are elaborated in Table 1 which shows the three general categories of testing practices (writing, development, and validation) and the general suggestions made in this column, but also summarizes the solutions offered for ways to implement those suggestions.

If even a few teachers begin to use a few of these suggestions, I have no doubt that their testing and therefore their teaching will improve. As a result, they will be better serving their students, themselves, and their institutions.

Table 1. ***Summary of Practices, Problems, and Solutions in Classroom Testing***

Practices	Problems	Solutions
Test Writing	Some teachers allow too little time for writing their test items (perhaps because they lack training in item writing)	*Create good quality items* by getting ahold of a good book on language testing and reading up on what good quality items are and how to write them; be sure to allot sufficient time by starting early.
	Tests sometimes seem like a disorganized hodge-podge of items	*Organize the items* by keeping items that are testing the same language point together; grouping items of the same format (e.g., multiple-choice, true-false, etc.); and keeping reading or listening items together with their passages.
	Students may find the organization of a test confusing, or worse, they may not understand what they need to do	*Provide clear headings and directions* by emphasizing headings (using bold, italics, etc.) and using them hierarchically; writing directions in students' mother tongue or in very simple/clear English; and using general and specific directions.
Development	Even with all of the above, other problems may remain (e.g., typos, spelling errors, etc.)	*Proofread the test* carefully yourself and get others to do so as well (including perhaps a colleague, former student, or even spouse) because another set of eyes can spot things you are too close to the test to see.
	Some teachers try to test their course objectives with too few items	*Use a sufficient number of items* by always writing 50% more good quality items than you think you will need; use common sense in deciding how many items to use while taking into account time constraints and the nature of the items.
	Some teachers fail to analyze and revise items even though they will use them again	*Examine the students' performances on the items* by listening to their questions/ comments during the test and revising; by considering the first administration a pilot test and performing item analysis (i.e., the *difference index* and *B index*) and revising.
Validation	Some teachers report the number of items correct and explain scores in terms of the bell curve	*Report the scores as percentages* and explain to students that the scores reveal how much they learned of the material taught in the course, rather than how the scores spread them out.
	Some teachers fail to consider & check if their score-based decisions are founded on unreliable information	*Check the reliability of the test items* in terms of sufficiency of information (the degree to which you have enough information to make consistent decisions) and calculate and interpret a K-R21 reliability coefficient.
	Some teachers fail to consider & check the validity of the scores on their tests	*Think about common sense validity issues* in terms of the degree to which the scores are measuring what you intended and the consequences/implications of your score uses by answering the six validity questions posed above.

References

Brown, J. D. (2003). Questions and answers about language testing statistics: Criterion-referenced item analysis (The difference index and B-index). *SHIKEN: The JALT Testing & Evaluation SIG Newsletter, 7*(3), 18-24). Accessed online September 29, 2013 at http://jalt.org/test/bro_18.htm

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

Brown, J. D. (2012). 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., & Hudson, T. (2002). *Criterion-referenced language testing.* Cambridge: Cambridge University.

Carr, N. T. (2011). *Designing and analyzing language tests.* Oxford: Oxford University.

Smith, J. K. (2003). Reconsidering reliability in classroom assessment and grading. *Educational Measurement: Issues and practice, 22*(4), 26-33.

Where to Submit Questions:

Please submit questions for this column to the following e-mail or snail-mail addresses:

brownj@hawaii.edu.

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

Your question can remain anonymous if you so desire.