

The Cornerstones of Testing

Language testing at any level is a highly complex undertaking that must be based on theory as well as practice. Although this book focuses on practical aspects of classroom testing, an understanding of the basic principles of larger-scale testing is essential. The nine guiding principles that govern good test design, development, and analysis are *usefulness, validity, reliability, practicality, washback, authenticity, transparency, and security*. Repeated references to these cornerstones of language testing will be made throughout this book.

Usefulness

For Bachman and Palmer (1996), the most important consideration in designing and developing a language test is the use for which it is intended: "Test usefulness provides a kind of metric by which we can evaluate not only the tests that we develop and use, but also all aspects of test development and use" (p. 17). Thus, *usefulness* is the most important quality or cornerstone of testing. Bachman and Palmer's model of test usefulness requires that any language test must be developed with a specific purpose, a particular group of test-takers, and a specific language use in mind.

Validity

The term *validity* refers to the extent to which a test measures what it purports to measure. In other words, *test what you teach and how you teach it!* Types of validity include content, construct, and face validity. For classroom teachers, *content validity* means that the test assesses the course content and outcomes using formats familiar to the students. *Construct validity* refers to the "fit" between the underlying theories and methodology of language learning and the type of assessment. For example, a communicative language learning approach must be matched by communicative language testing. *Face validity* means that the test looks as though it measures what it is supposed to measure. This is an important factor for both students and administrators. Moreover, a professional-looking exam has more credibility with students and administrators than a sloppy one.

It is important to be clear about what we want to assess and then be certain that we are assessing that material and not something else. Making sure that clear assessment objectives are met is of primary importance in achieving test validity. The best way to ensure validity is to produce tests to specifications. See Chapter 1 regarding the use of specifications.

Reliability

Reliability refers to the consistency of test scores, which simply means that a test would offer similar results if it were given at another time. For example, if the same test were to be administered to the same group of students at two different times in two different settings, it should not make any difference to the test-taker whether he or she takes the test on one occasion and in one setting or the other. Similarly, if we develop two forms of a test that are intended to be used interchangeably, it should not make any difference to the test-taker which form or version of the test he or she takes. The student should obtain approximately the same score on either form or version of the test. Versions of exams that are not equivalent can be a threat to reliability, the use of specifications is strongly recommended; developing all versions of a test according to specifications can ensure equivalency across the versions.

Three important factors affect test reliability. Test factors such as the formats and content of the questions and the time given for students to take the exam must be consistent. For example, testing research shows that longer exams produce more reliable results than brief quizzes (Bachman, 1990, p. 220). In general, the more items on a test, the more reliable it is considered to be because teachers have more samples of students' language ability. Administrative factors are also important for reliability. These include the classroom setting (lighting, seating arrangements, acoustics, lack of intrusive noise, etc.) and how the teacher manages the administration of the exam. Affective factors in the response of individual students can also affect reliability, as can fatigue, personality type, and learning style. Test anxiety can be allayed by coaching students in good test-taking strategies.

A fundamental concern in the development and use of language tests is to identify potential sources of error in a given measure of language ability and to minimize the effect of these factors on test reliability. Henning (1987) describes these threats to test reliability.

- **Fluctuations in the Learner.** A variety of changes may take place within the learner that may change a learner's true score from test to test. Examples of this type of change might be additional learning or forgetting. Influences such as fatigue, sickness, emotional problems, and practice effect may cause the learner's score to deviate from the score that reflects his or her actual ability. Practice effect means that a student's score could improve because he or she has taken the test so many times that the content is familiar.

- **Fluctuations in Scoring.** Subjectivity in scoring or mechanical errors in the scoring process may introduce error into scores and affect the reliability of the test's results. These kinds of errors usually occur within (intra-rater) or between (inter-rater) the raters themselves.
- **Fluctuations in Test Administration.** Inconsistent administrative procedures and testing conditions will reduce test reliability. This problem is most common in institutions where different groups of students are tested in different locations on different days.

Reliability is an essential quality of test scores because unless test scores are relatively consistent, they cannot provide us with information about the abilities we want to measure. A common theme in the assessment literature is the idea that reliability and validity are closely interlocked. While reliability focuses on the empirical aspects of the measurement process, validity focuses on the theoretical aspects and interweaves these concepts with the empirical ones (Davies et al., 1999, p. 169). For this reason it is easier to assess reliability than validity.

Practicality

Another important feature of a good test is practicality. Classroom teachers know all too well the importance of familiar practical issues, but they need to think of how practical matters relate to testing. For example, a good classroom test should be "teacher friendly." A teacher should be able to develop, administer, and mark it within the available time and with available resources. Classroom tests are only valuable to students when they are returned promptly and when the feedback from assessment is understood by the student. In this way, students can benefit from the test-taking process. Practical issues include the cost of test development and maintenance, adequate time (for development and test length), resources (everything from computer access, copying facilities, and AV equipment to storage space), ease of marking, availability of suitable/trained graders, and administrative logistics. For example, teachers know that ideally it would be good to test speaking one-on-one for up to ten minutes per student. However, for a class of 25 students, this could take four hours. In addition, what would the teachers do with the other 24 students during the testing?

Washback ^{test impact}

Washback refers to the effect of testing on teaching and learning. Washback is generally said to be positive or negative. Unfortunately, students and teachers

tend to think of the negative effects of testing such as “test-driven” curricula and only studying and learning “what they need to know for the test.” In contrast, positive washback, or what we prefer to call *guided washback*, benefits teachers, students, and administrators because it assumes that testing and curriculum design are both based on clear course outcomes that are known to both students and teachers/testers. If students perceive that tests are markers of their progress toward achieving these outcomes, they have a sense of accomplishment.

Authenticity

Language learners are motivated to perform when they are faced with tasks that reflect real-world situations and contexts. Good testing or assessment strives to use formats and tasks that mirror the types of situations in which students would authentically use the target language. Whenever possible, teachers should attempt to use authentic materials in testing language skills. For K-12 teachers of content courses, the use of authentic materials at the appropriate language level provides additional exposure to concepts and vocabulary as students will encounter them in real-life situations.

Transparency

Transparency refers to the availability of clear, accurate information to students about testing. Such information should include outcomes to be evaluated, formats used, weighting of items and sections, time allowed to complete the test, and grading criteria. Transparency dispels the myths and mysteries surrounding testing and the sometimes seemingly adversarial relationship between learning and assessment. Transparency makes students part of the testing process.

Security

Most teachers feel that security is an issue only in large-scale, high-stakes testing. However, security is part of both reliability and validity for all tests. If a teacher invests time and energy in developing good tests that accurately reflect the course outcomes, then it is desirable to be able to recycle the test materials. Recycling is especially important if analyses show that the items, distractors, and test sections are valid and discriminating. In some parts of the world, cultural attitudes toward “collaborative test-taking” are a threat to test security and thus to reliability and validity. As a result, there is a trade-off between letting tests into the public domain and giving students adequate information about tests.

Ten Things to Remember

1. Test what has been taught and how it has been taught.

This is the basic concept of content validity. In achievement testing, it is important to only test students on what has been covered in class and to do this through formats and techniques they are familiar with.

2. Set tasks in context whenever possible.

This is the basic concept of authenticity. Authenticity is just as important in language testing as it is in language teaching. Whenever possible, develop assessment tasks that mirror purposeful real-life situations.

3. Choose formats that are authentic for tasks and skills.

Although challenging at times, it is better to select formats and techniques that are purposeful and relevant to real-life contexts.

4. Specify the material to be tested.

This is the basic concept of transparency. It is crucial that students have information about how they will be assessed and have access to the criteria on which they will be assessed. This transparency will lower students' test anxiety.

5. Acquaint students with techniques and formats prior to testing.

Students should never be exposed to a new format or technique in a testing situation. Doing so could affect the reliability of your test/assessment. Don't avoid new formats; just introduce them to your classes in a low-stress environment outside the testing situation.

6. Administer the test in uniform, non-distracting conditions.

Another threat to the reliability of your test is the way in which you administer the assessment. Make sure your testing conditions and procedures are consistent among different groups of students.

7. Provide timely feedback.

Feedback is of no value if it arrives in the students' hands too late to do anything with it. Provide feedback to students in a timely manner. Give easily scored objective tests back during the next class. Aim to return subjective tests that involve more grading within three class periods.

8. Reflect on the exam without delay.

Often teachers are too tired after marking the exam to do anything else. Don't shortchange the last step—that of reflection. Remember, all stakeholders in the exam process (that includes you, the teacher) must benefit from the exam.

9. Make changes based on analyses and feedback from colleagues and students.

An important part of the reflection phase is the opportunity to revise the exam when it is still fresh in your mind. This important step will save you time later in the process.

10. Employ multiple measures assessment in your classes.

Use a variety of types of assessment to determine the language abilities of your students. No one type of assessment can give you all the information you need to accurately assess your students.