Introduction to Issues in Language Assessment and Terminology

In today's language classrooms, the term *assessment* usually evokes images of an end-of-course paper-and-pencil test designed to tell both teachers and students how much material the student doesn't know or hasn't yet mastered. However, assessment is much more than tests. Assessment includes a broad range of activities and tasks that teachers use to evaluate student progress and growth on a daily basis.

Consider a day in the life of Ms. Wright, a typical experienced ESL teacher in a large urban secondary school in Florida. In addition to her many administrative responsibilities, she engages in a wide range of assessment-related tasks on a daily basis. It is now May, two weeks before the end of the school year. Today, Ms. Wright did the following in her classroom:

- graded and analyzed yesterday's quiz on the irregular past tense
- decided on topics for tomorrow's review session
- administered a placement test to a new student to gauge the student's writing ability
- met with the principal to discuss the upcoming statewide exam
- checked her continuous assessment records to choose students to observe for speaking today
- improvised a review when it was clear that students were confused about yesterday's vocabulary lesson
- made arrangements to offer remediation to students who did poorly on last week's reading practice exam
- after reviewing the final exam that came with the textbook, decided to revise questions to suit class focus and coverage
- graded students' first drafts of a travel webquest using checklists distributed to students at the start of the project

Each of these tasks was based on a decision Ms. Wright made about her students or her class as a whole. Teachers assess their students in a number of ways and for a variety of purposes because they need to make decisions about their classrooms and their teaching. Some of these decisions are made on the spot, such as the improvised review. Others, like preparing the final exam, entail long-term planning.

Placing students in the right level of classroom instruction is an essential purpose of assessment. Normally, new students are given placement exams at the beginning of the school year, but some new students arrive throughout the year. By assigning a new student a writing task to gauge her writing ability, Ms. Wright tried to ensure that the student would benefit from instruction at the appropriate level for the remaining weeks of the school year.

Some of the decisions Ms. Wright made today had to do with **diagnosing student problems.** One of a teacher's main aims is to identify students' strengths and weaknesses with a view to carrying out revision or remedial activities. By making arrangements to offer remediation to students who did poorly on last week's reading exam, she was engaging in a form of **diagnostic assessment.**

Much of what teachers do today in language classrooms is to **find out about the language proficiency of their students.** In preparing her students to take the Florida Comprehensive Assessment Test (FCAT), Ms. Wright was determining whether her students have sufficient language proficiency to complete the exam effectively and meet national benchmarks.

Other activities were carried out with the aim of **evaluating academic performance.** In fact, a lot of teacher time is spent gathering information that will help teachers make decisions about their students' achievement regarding course goals and mastery of course content. Ms. Wright uses multiple measures such as quizzes, tests, projects, and continuous assessment to monitor her students' academic performance. To assign speaking grades to her students, she had to select four or five students per day for her continuous assessment records. These daily speaking scores will later be averaged together with her students' formal oral interview results for their final speaking grades.

Many of her classroom assessment activities concerned **instructional decision-making.** In deciding which material to present next or what to revise, Ms. Wright was making decisions about her language classroom. When she prepares her lesson plans, she consults the syllabus and the course objectives, but she also makes adjustments to suit the immediate needs of her students.

Some of the assessment activities that teachers participate in are for **accountability purposes.** Teachers must provide educational authorities with evidence that their intended learning outcomes have been achieved. Ms. Wright understands that her assessment decisions impact her students, their families, her school administration, and the community in which she works.

Evaluation, Assessment, and Testing

To help teachers make effective use of evaluation, assessment, and testing procedures in the foreign/second (F/SL) language classroom, it is necessary to clarify what these concepts are and explain how they differ from one another.

The term *evaluation* is all-inclusive and is the widest basis for collecting information in education. According to Brindley (1989), evaluation is "conceptualized as broader in scope, and concerned with the overall program" (p. 3). Evaluation involves looking at all factors that influence the learning process, i.e., syllabus objectives, course design, and materials (Harris & McCann, 1994). Evaluation goes beyond student achievement and language assessment to consider all aspects of teaching and learning and to look at how educational decisions can be informed by the results of alternative forms of assessment (Genessee, 2001).

Assessment is part of evaluation because it is concerned with the student and with what the student does (Brindley, 1989). *Assessment* refers to a variety of ways of collecting information on a learner's language ability or achievement. Although *testing* and *assessment* are often used interchangeably, *assessment* is an umbrella term for all types of measures used to evaluate student progress. *Tests* are a subcategory of assessment. A *test* is a formal, systematic (usually paper-and-pencil) procedure used to gather information about students' behavior.

In summary, *evaluation* includes the whole course or program, and information is collected from many sources, including the learner. While *assessment* is related to the learner and his or her achievements, *testing* is part of assessment, and it measures learner achievement.

Categorizing Assessment Tasks

Different types of tests are administered for different purposes and used at different stages of the course to gather information about students. You as a language teacher have the responsibility of deciding on the best option for your particular group of students in your particular teaching context. It is useful to categorize assessments by type, purpose, or place within the teaching/learning process or timing.

Types of Tests

The most common use of language tests is to identify strengths and weaknesses in students' abilities. For example, through testing we might discover that a student has excellent oral language abilities but a relatively low level of reading comprehension. Information gleaned from tests also assists us in deciding who should be allowed to participate in a particular course or program area. Another common use of tests is to provide information about the effectiveness of programs of instruction.

Placement Tests

Placement tests assess students' level of language ability so they can be placed in an appropriate course or class. This type of test indicates the level at which a student will learn most effectively. The primary aim is to create groups of learners that are homogeneous in level. In designing a placement test, the test developer may base the test content either on a theory of general language proficiency or on learning objectives of the curriculum. Institutions may choose to use a well-established proficiency test such as the TOEFL®, IELTS™, or MELAB exam and link it to curricular benchmarks. Alternatively, some placement tests are based on aspects of the syllabus taught at the institution concerned (Alderson, Clapham, & Wall, 1995).

At some institutions, students are placed according to their overall rank in the test results combined from all skills. At other schools and colleges, students are placed according to their level in each skill area. Additionally, placement test scores are used to determine if a student needs further instruction in the language or could matriculate directly into an academic program without taking preparatory language courses.

Aptitude Tests

An *aptitude test* measures capacity or general ability to learn a foreign or second language. Although not commonly used these days, two examples deserve mention: the Modern Language Aptitude Test (MLAT) developed by Carroll and Sapon in 1958 and the Pimsleur Language Aptitude Battery (PLAB) developed by Pimsleur in 1966 (Brown, H.D., 2004). These are used primarily in deciding to sponsor a person for special training based on language aptitude.

Diagnostic Tests

Diagnostic tests identify language areas in which a student needs further help. Harris and McCann (1994) point out that where "other types of tests are based on success, diagnostic tests are based on failure" (p. 29). The information gained from diagnostic tests is crucial for further course activities and providing students with remediation. Because diagnostic tests are difficult to write, placement tests often serve a dual function of both placement and diagnosis (Harris & McCann, 1994; Davies et al., 1999).

Progress Tests

Progress tests measure the progress that students are making toward defined course or program goals. They are administered at various stages throughout a language course to determine what students have learned, usually after certain segments of instruction have been completed. Progress tests are generally teacher produced and narrower in focus than achievement tests because they cover less material and assess fewer objectives.

Achievement Tests

Achievement tests are similar to progress tests in that they determine what a student has learned with regard to stated course outcomes. They are usually administered at mid- and end-point of the semester or academic year. The content of achievement tests is generally based on the specific course content or on the course objectives. Achievement tests are often cumulative, covering material drawn from an entire course or semester.

Proficiency Tests

Proficiency tests, on the other hand, are not based on a particular curriculum or language program. They assess the overall language ability of students at varying levels. They may also tell us how capable a person is in a particular language skill area (e.g., reading). In other words, proficiency tests describe what students are capable of doing in a language.

Proficiency tests are typically developed by external bodies such as examination boards like Educational Testing Services (ETS), the College Board, or Cambridge ESOL. Some proficiency tests have been standardized for international use, such as the TOEFL®, which measures the English language proficiency of foreign college students who wish to study in North American universities or the IELTS™, which is intended for those who wish to study in the United Kingdom or Australia (Davies et al., 1999). Increasingly, North American universities are accepting IELTS™ as a measure of English language proficiency.

Additional Ways of Labeling Tests

Objective versus Subjective Tests

Sometimes tests are distinguished by the manner in which they are scored. An *objective test* is scored by comparing a student's responses with an established set of acceptable/correct responses on an answer key. With objectively scored tests, the scorer does not require particular knowledge or training in the examined area. In contrast, a *subjective test,* such as writing an essay, requires scoring by opinion or personal judgment so the human element is very important.

Testing formats associated with objective tests are multiple choice questions (MCQs), True/False/Not Given (T/F/Ns), and matching. Objectively scored tests are ideal for computer scanning. Examples of subjectively scored tests are essay tests, interviews, or comprehension questions. Even experienced scorers or markers need moderated training sessions to ensure inter-rater reliability.

Criterion-Referenced versus Norm-Referenced or Standardized Tests

Criterion-referenced tests (CRTs) are usually developed to measure mastery of well-defined instructional objectives specific to a particular course or program. Their purpose is to measure how much learning has occurred. Student performance is compared only to the amount or percentage of material learned (Brown, J.D., 2005).

True CRTs are devised before instruction is designed so that the test will match the teaching objectives. This lessens the possibility that teachers will "teach to the test." The criterion or cut-off score is set in advance. Student achievement is measured with respect to the degree of learning or mastery of the pre-specified content. A primary concern of a CRT is that it be sensitive to different ability levels.

Norm-referenced tests (NRT) or standardized tests differ from criterion-referenced tests in a number of ways. NRTs are designed to measure global language abilities. Students' scores are interpreted relative to all other students who take the exam. The purpose of an NRT is to spread students out along a continuum of scores so that those with low abilities in a certain skill are at one end of the normal distribution and those with high scores are at the other end, with the majority of the students falling between the extremes (Brown, J.D., 2005, p. 2).

By definition, an NRT must have been previously administered to a large sample of people from the target population. Acceptable standards of achievement are determined after the test has been developed and administered. Test results are interpreted with reference to the performance of a given group or norm. The *norm* is typically a large group of students who are similar to the individuals for whom the test is designed.

Summative versus Formative

Tests or tasks administered at the end of the course to determine if students have achieved the objectives set out in the curriculum are called *summative assessments.* They are often used to decide which students move on to a higher level (Harris & McCann, 1994). *Formative assessments,* however, are carried out with the aim of using the results to improve instruction, so they are given during a course and feedback is provided to students.

High-Stakes versus Low-Stakes Tests

High-stakes tests are those in which the results are likely to have a major impact on the lives of large numbers of individuals or on large programs. For example, the TOEFL® is high stakes in that admission to a university program is often contingent on receiving a sufficient language proficiency score.

Low-stakes tests are those in which the results have a relatively minor impact on the lives of the individual or on small programs. In-class progress tests or short quizzes are examples of low-stakes tests.

Traditional versus Alternative Assessment

One useful way of understanding alternative assessment is to contrast it with traditional testing. *Alternative assessment* asks students to show what they can do; students are evaluated on what they integrate and produce rather than on what they are able to recall and reproduce (Huerta-Macias, 1995). Competency-based assessment demonstrates what students can actually *do* with English. Alternative assessment differs from traditional testing in that it:

- does not intrude on regular classroom activities
- reflects the curriculum actually being implemented in the classroom
- provides information on the strengths and weaknesses of each individual student
- provides multiple indices that can be used to gauge student progress
- is more multiculturally sensitive and free of the linguistic and cultural biases found in traditional testing (Huerta-Macias, 1995).

Types of Alternative Assessment

Several types of alternative assessment can be used with great success in today's language classrooms:

- Self-assessment
- Portfolio assessment
- Student-designed tests
- Learner-centered assessment
- Projects
- Presentations

Specific types of alternative assessment will be discussed in the skills chapters.

This chart summarizes common types of language assessment.

Table 1: Common Types of Language Assessment

Informal	Formal
Classroom, "low-stakes"	Standardized, "high-stakes"
Criterion-referenced	Norm-referenced
Achievement	Proficiency
Direct	Indirect
Subjective	Objective
Formative	Summative
Alternative, authentic	Traditional tests

Because language performance depends heavily on the purpose for language use and the context in which it is used, it makes sense to provide students with assessment opportunities that reflect these practices. *Our assessment practices must reflect the importance of using language both in and out of the language classroom.*

It is also important to note that most testers today recommend that teachers use *multiple measures assessment.* Multiple measures assessment comes from the belief that no single measure of language assessment is enough to tell us all we need to know about our students' language abilities. That is, we must employ a mixture of all the assessment types previously mentioned to obtain an accurate reading of our students' progress and level of language proficiency.

Test Purpose

One of the most important first tasks of any test writer is to determine the purpose of the test. Defining the purpose aids in selection of the right type of test. This table shows the purpose of many of the common test types.

Table 2: Common Test Types

Test Type	Main Purpose
Placement tests	Place students at appropriate level of instruction within program
Diagnostic tests	Identify students' strengths and weaknesses for remediation
Progress tests or in-course tasks	Provide information about mastery or difficulty with course materials
Achievement tests	Provide information about students' attainment of course outcomes at end of course or within the program
Standardized tests	Provide measure of students' proficiency using international benchmarks

Timing of the Test

Tests are commonly categorized by the point in the instructional period at which they occur. Aptitude, admissions, and general proficiency tests often take place before or outside of the program; placement and diagnostic tests often occur at the start of a program. Progress and achievement tests take place during the course of instruction and promotion, while mastery or certification tests occur at the end of a course of study or program.

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

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 constrast, 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.

Extension Activity

Cornerstones Case Study

Read this case study about Mr. Knott, a colleague of Ms. Wright's, and try to spot the cornerstones violations. What could be done to solve these problems?

Background Information

Mr. Knott is a high school ESL and Spanish teacher. His current teaching load is two ESL classes. His students come from many language backgrounds and cultures. In his classes, he uses an integrated-skills textbook that espouses a communicative methodology.

His Test

Mr. Knott firmly believes in the KISS philosophy of "keeping it short and simple." Most recently he has covered modal verbs in his classes. He decides to give his students only one question to test their knowledge about modal verbs: "Write a 300-word essay on the meanings of modal verbs and their stylistic uses. Give examples and be specific." Because he was short of time, he distributed a handwritten prompt on unlined paper. Incidentally, he gave this same test last year.

Information Given to Students

To keep students on their toes and to increase attendance, he told them that the test could occur anytime during the week. Of his two classes, Mr. Knott has a preference for his morning class because they are much more well behaved and hard working so he hinted during the class that modal verbs might be the focus of the test. His afternoon class received no information on the topic of the test.

Test Administration Procedures

Mr. Knott administered his test to his afternoon class on Monday and to his morning class on Thursday. Always wanting to practice his Spanish, he clarified the directions for his Spanish-speaking students in Spanish. During the Monday administration, his test was interrupted by a fire drill. Since this was the first time a fire drill had happened, he did not have any back-up plan for collecting test papers. Consequently, some students took their papers with them. In the confusion, several test papers were mislaid.

Grading Procedures

Mr. Knott didn't tell his students when to expect their results. Due to his busy schedule, he graded tests over several days during the next week. Students finally got their tests back ten days later. Because the test grades were extremely low, Mr. Knott added ten points to everyone's paper to achieve a good curve.

Post-Exam Follow-Up Procedures

Mr. Knott entered grades in his grade book but didn't annotate or analyze them. Although Mr. Knott announced in class that the exam was worth 15 percent of the students' grade, he downgraded it to five percent. Next year he plans to recycle the same test but will require students to write 400 words.

What's wrong with Mr. Knott's testing procedures? Your chart should look something like this.

Cornerstone Violation	Mr. Knott's Problem	Possible Solution
	Construct validity violation: • He espouses a communicative language teaching philosophy but gives a test that is not communicative.	
	Authenticity violation: • Writing about verb functions is not an authentic use of language.	Mr. Knott should have chosen tasks that required students to use modal verbs in real-life situations.
	Practicality violation: • He was short of time and distributed a handwritten test.	Mr. Knott probably waited until the last minute and threw something together in panic mode.
	Face validity violation: • He distributed a hand-written prompt on unlined paper.	Tests must have a professional look.
	Security violation: • He gave the same test last year, and it's probably in the public domain.	If a test was administered verbatim the previous year, there is a strong probability that students already have access to it. Teachers should make every effort to produce parallel forms of tests that are secure.

Cornerstone Violation	Mr. Knott's Problem	Possible Solution
Information Given to Students	**Transparency violation:** • He preferred one class over another (potential bias) and gave them more information about the test.	Mr. Knott needs to provide the same type and amount of information to all students.
Test Administration Procedures	**Security violation:** • He administered the same test to both classes three days apart. • Some students took their papers outside during the fire drill. • Some students lost their papers. **Reliability/transparency violation:** • His Spanish-speaking students got directions in Spanish.	When administering the same test to different classes, an effort should be made to administer the tests close together so as to prevent test leaks. Mr. Knott should have disallowed this test due to security breaches. The same type and amount of information should be given to all students.
Grading Procedures	**Transparency violation:** • Students didn't know when to expect their results. **Reliability violation:** • He graded test papers over the course of a week (i.e., there was potential for intra-rater reliability problems). **Washback violation:** • Students got their papers back ten days later so there was no chance for remediation.	Teachers should return test papers to students no longer than three class periods after the test was administered. It would have been better to grade all the papers in a shorter period of time to ensure a similar internal standard of marking. As students were already into the next set of objectives, they had no opportunity to practice material they did poorly on. Teachers should always return papers in a timely manner and review topics that proved problematic for students.

Cornerstone Violation	Mr. Knott's Problem	Possible Solution
Post-Exam Follow-Up Procedures	**Security violation:** • He plans to recycle the test yet again.	Only good tests should be recycled. Mr. Knott's students didn't do so well on this test, and he had to curve the grades. This should tell Mr. Knott that the test needs to be retired or seriously revised.