Introduction

Testing in general strikes a deep emotional chord in people. Those who enjoy competition are up for the challenge, but most of us would rather forgo the ordeal. Language testing is particularly daunting because it does not just attempt to assess our knowledge about a specific subject but assesses how effectively, as test-takers, we are capable of communicating with the rest of the world.

At the same time that a test is making examinees fearful, teachers and test developers must worry about basic issues of fairness, utility, and the power tests hold over other people's futures. Sometimes high stakes are involved: passing a course, qualifying for study abroad, admission to a degree program, or having access to citizenship or employment. To sympathetic EFL/ESL teachers, anxious to see their students develop self-confidence and reluctant to crush their efforts, assessing their students' skills can be painful. They seek to empower and not judge.

With all of this emotion and trepidation stirred up by assessment and how it is performed, is it any wonder that myths should develop?

This book, *Assessment Myths: Applying Second Language Research to Classroom Teaching,* follows others in the University of Michigan Press series on myths in language learning and teaching. This volume has adopted the same framework as previous Myths books to illuminate important issues that impact language teaching and learning. Each chapter presents a myth, reviews research that qualifies or debates the myth, and concludes with suggestions for applying the research. It is our hope that this book will provide interesting conversations as well as valuable content for courses in language teaching and assessment.

How do we define language assessment? How did we decide which eight myths to include? What are some key terms in language assessment? These fundamental issues will be introduced first.

What Is Language Assessment?

In language learning, evaluation should be an integral part of the teachers' decision-making, their students' self-reflection, and the long-term instructional operations. Assessment takes different forms and serves many purposes in instructional contexts. These purposes can range from a student answering a teacher's question to an applicant taking an entrance exam. In addition, the uses of language assessment are broad and life-altering, from informing teaching to fulfilling citizenship requirements. Language teachers and programs use assessment to make many decisions—diagnostic, placement, exiting, and hiring, to name just a few.

Given their ubiquitous and powerful status, language assessments should be considered carefully and developed thoughtfully. We look to assessment to provide information about language abilities. This goal is often more complicated than it first appears. For example, interviewing language learners or asking them to write on a given topic provides a seemingly transparent picture of their ability to use language. However, their responses can be affected by many issues other than language such as the test-taker's familiarity with the topic, prior experience with tests, level of anxiety, and clarity of the instructions for the task. When assessing learners' listening skills, we cannot actually know what they comprehend without asking them to use other skills, such as reading test questions or writing a response. Despite or maybe because of these complications, the field of language assessment has grown over the years and continues to seek answers and best practices for these dilemmas. This book provides an overview of research, theory, and application on many of the issues in language testing.

Developing an assessment is difficult for the reasons just explained. Even a well-developed test can cause challenges for teachers if it does not measure the skills to be evaluated. The purpose of the test should be a primary concern in selecting what to use. For example, an end-of-term exam should be designed to include the skills and content covered in the course. If a test's design is not aligned with its purpose, then the test results may not be meaningful. Such misalignment is not uncommon and will be a recurring theme throughout this book because that alignment is the key to valid, ethical, and fair assessment. This issue is important in both large-scale standardized tests and classroom assessments. As teachers, we need to ask ourselves regularly, "Why am I giving this test?" and "What do I hope to learn about my students/class/teaching?" The answers to these questions should help identify useful tests and remind us that testing should be for learning and not for unknown or, even worse, punitive purposes.

Assessments come in many shapes and forms. Probably the first thing that comes to mind involves performances on tasks or questions to be answered in an allotted amount of time. Perhaps the most familiar form is a paper-based test that includes multiple choice, fill-in-the-blank, and/or sentence completion questions. More recently language testing has included performance-based assessments, such as participating in an interview or writing a short essay. In classroom assessment, the **process** can be evaluated along with the **product;** the preparation of portfolios requires students to first collect samples of their work over a period of time and then reflect on their choices. This depth of information about students' progress during the semester enlightens end-of-course decisions and helps with future placement. Each of these formats has advantages and shortcomings that should be considered when developing language assessments.

Why These Eight Myths?

The power that tests have in affecting people's lives is one likely reason for the many myths that exist about assessment. Test-takers and students create theories about how to get a high score, while teachers or administrators ponder how to interpret test scores. The nature of myths is that they are not entirely false. Anthropologists study myths in cultures because they reflect worldviews, values, and explanations for inexplicable phenomena. The assessment myths explored in this book also have reasons for existing and, in some cases, contain some element of truth. In each chapter we will dissect the myth to illuminate the complexities of the topic as well as what can make it problematic for classroom teachers.

As assessment specialists and long-time language teachers, we have encountered many myths, and at times have held some myths of our own. We used these experiences to select myths for this book. We narrowed our list to eight myths by asking ourselves three questions: (1) Does this myth address a critical issue in language assessment? (2) Are there theories and research that delve into the truth behind this myth? and (3) Is this myth important to teachers? The answers left us with these myths:

Myth 1: Assessment is just writing tests and using statistics.

Myth 2: A comprehensive final exam is the best way to evaluate students.

Myth 3: Scores on performance assessments are preferable because of their accuracy and authenticity.

Myth 4: Multiple choice questions are inaccurate measures of language but are easy to write.

Myth 5: We should test only one skill at a time.

Myth 6: A test's validity can be determined by looking at it.

Myth 7: Issues of fairness are not a concern with standardized testing.

Myth 8: Teachers should not be involved in preparing students for tests.

The order of these eight myths is deliberate. The first myth builds the foundation for the later myths by expanding the definition of what it means to know about assessment. Myths 2 through 5 discuss how the various kinds of assessment can be used effectively, such as classroom- and performance-based tests, multiple choice questions, and integrated skills assessments. The last three myths (6, 7, and 8) focus on theoretical and philosophical issues that have practical implications. They renew emphasis on the importance of topics in the earlier chapters. We hope that, by addressing eight beliefs about language assessment, this book will provide helpful insights into language assessment and instruction.

Key Terms in Language Testing

Throughout this book, several terms are important to general understanding of testing as well as to specific myths. This section explains a number of key ideas in broad terms because most of them will be developed and illustrated in later chapters. These terms are grouped in three categories: testing purposes, paired terms, and development and use of assessment in educational contexts.

TESTING PURPOSES

Each type of assessment fulfills different needs of teachers and language programs. It is critical to recognize the purpose of a test to appropriately develop or select a measure to fit one or more of these objectives:

- **Diagnostic:** This type of assessment is used to determine students' specific strengths and weaknesses during an instructional program. It helps teachers plan a path of improvement for each learner. These tests need to be aligned with the course curriculum and goals. Teachers can use the test results of the whole class to make decisions as to which aspects of language they will focus on and what are less important concerns for their students.
- **Placement:** A placement exam helps identify a student's level of language competency to determine which

course or level is appropriate. Placement testing is common in language programs that offer different levels of a language course. It is challenging because test-takers will have a range of abilities in the target language, such as having good reading skills but weak speaking proficiency. Combining students with different proficiency profiles into a course level so that all of them are well served is a tricky business. Teachers are often involved in placement testing, serving as proctors, raters, or even decision-makers.

- **Achievement:** These tests are developed to assess students' learning of coursework or other curriculum materials. They are closely aligned to course or program objectives and are often used at the end of a learning sequence. Achievement tests are among the most relevant for classroom teachers.
- **Proficiency:** Such general tests of language ability are often developed for large-scale testing without planned alignment to specific learning goals or language programs. Specific proficiency examinations, such as TOEFL® and IELTS®, serve as an admission requirement for non-native English speakers at many colleges and universities that use English as the medium for instruction. Government agencies, licensing bodies, businesses, and other institutions may also mandate proficiency tests. Teachers usually encounter proficiency tests when the administration of their institution requires a specific score from applicants for entry.

PAIRED TERMS

Quite a few terms in the field of assessment are paired because they are related and, in many cases, because they are contrasting concepts. Learning these terms together can be useful for comprehension and can illuminate important controversies in language testing.

- **Formative assessment** versus **summative assessment:** Formative assessment is used routinely in classrooms to reveal whether the students are mastering the learning objectives (see Myth 2). It generates information to gauge progress and helps improve learning and teaching throughout the course. Summative assessment provides information on learners' achievement at the end of a course or program.
- **High-stakes/standardized testing** versus **classroom-based assessment:** Any test that is given in the same manner to all test-takers can be considered a standardized test. We tend to contrast assessment used for high-stakes decisions developed by large publishers or testing organizations with assessment that is used or developed by individual teachers for classroom purposes with their particular students.
- **Selected response** versus **constructed response:** This distinction is about different kinds of test tasks. Selected-response test items require students to choose the correct answer, for example, in multiple choice questions or matching tasks. Constructed-response tasks ask test-takers to create or develop a response, either in a sentence or in longer spoken or written responses. Selected-response items are frequently used in large-scale testing for efficiency; however, in classroom assessments, constructed-response tasks are likely to be more meaningful, depending on course goals.
- **Performance assessment** versus **objective assessment:** Performance assessments include essay-writing or oral interviews, while objective assessments are generally selected-response items. The primary difference is in scoring: although some responses can be scored objectively when there is a right or wrong answer, performance assessment requires rating, which makes subjectivity possible and therefore complicates scoring.

If teachers use a performance assessment, such as an interview or essay, they need to be careful about consistency and fairness in how they score students' work.

- **Independent** versus **integrated skills assessment:** This pairing is about the kind of language being assessed. When reading, writing, listening, or speaking as an individual skill is the main focus of a test and the score will reflect ability in that particular skill, then it is **independent skills** assessment. On the other hand, if skills are combined, such as reading with writing or listening with speaking, then it is called **integrated skills** assessment. Choosing between these two approaches depends on the purpose of the assessment and the kind of language students will need to use in the future. If they need to combine skills, then integrated assessment is a good choice.
- **Validity** and **reliability:** Validity has to do with the confidence we have in our interpretations of test scores, while reliability is the consistency of those scores. Validity is important in deciding what meaning can be placed on the test results. For example, a test of speaking proficiency that requires test-takers to read and circle answers to multiple choice questions wouldn't have much validity. Reliability concerns the degree to which a test produces stable, consistent results. Factors affecting the reliability of a test include its length, whether it's scored objectively or subjectively, who the students are who are taking it, and what types of questions are included. Both of these concepts are important to ensure our assessments are of good quality and allow us to interpret the scores meaningfully. They are related to issues of fairness and accuracy, which are as important in classroom assessments as they are in standardized tests.

Development and Use of Assessment in Educational Contexts

The last group of terms are those commonly used in educational contexts to talk about the development and use of assessments.

- **Rubrics/rating scales:** Scoring scales are used to evaluate performances. They include criteria for a performance, descriptors of those criteria, and a scheme to arrive at a score. They can be used to improve the efficiency, fairness, and clarity of scores as well as to provide useful feedback on students' strengths and weaknesses.
- **Item analysis:** These procedures are for checking that test questions are at the right level of difficulty and that they distinguish test-takers appropriately. For example, if students who have a low overall score on a test get a certain item correct more often than higher-scoring students, this item needs to be rewritten. If a certain item is answered incorrectly by all test-takers, it might also be one that should be removed or rewritten. Item analysis also includes checking for biased questions and other problems that can occur in item writing. This process is done systematically on large-scale tests that use selected-response items and involves statistical analysis.
- **Score interpretation:** It is rare in language testing to assess language with the precision of measurement devices such as thermometers and rulers. Nevertheless, language assessment usually results in a number (score) that is used to reflect or interpret something about a student's language ability. While numbers are needed to provide students, teachers, and other stakeholders with a scale on which to evaluate improvement or general proficiency, it is important to remember that the number (score) itself does **not** carry meaning about a stu-

dent's ability. Instead, the way we interpret that score is what gives it meaning. For example, 80 out of 100 only gives us a ratio or percentage, but in the context of a test, it could mean a student should be placed in a high-intermediate class, or it could mean that the student's English is good enough for academic coursework. It could also mean that the student has been successful in meeting course goals or that a pre-service teacher needs to improve his or her language skills before qualifying to teach English.

- **Construct:** A construct is a theoretical model of the underlying ability we are trying to assess. Constructs can be based on research or they may be based on a curriculum or syllabus. For example, the construct for second language reading might be defined with two major components: fluency and comprehension, which may be further subdivided into the skills and strategies of vocabulary recognition, inference-making, prediction, or first-language transfer. Defining the construct should be the first step in developing an assessment, and it must be aligned with the test's purpose.
- **Test washback:** Washback is the impact of a particular test on teachers, students, classrooms, and the outside world. Washback can be positive or negative. Negative washback occurs when a test undermines teaching and detracts from quality learning. Conversely, positive washback comes about when a test is well aligned with a course, thus supporting and motivating teachers and students.
- **Bias:** In assessments, when an individual or group of test-takers has certain characteristics that are not part of the construct being assessed and that give an unfair advantage, we say the test is biased in their favor. Or, the test may be biased to the detriment of some students. An example of bias might be when language assessment

> shows performance patterns that follow gender lines (female students scoring higher than male students or vice versa). In such cases, the assessment needs careful scrutiny to ensure that test items or tasks are not biased and did not cause such a pattern.

Throughout this book's exploration of myths in language testing, we hope readers will gain insight in the field as well as useful ideas for their teaching. Assessment holds a permanent position in teaching and learning, and the more teachers can leverage it to support their students and their classroom practice, the better. We wrote the chapters of this book with this purpose in mind.