20
Large-Scale Second Language Assessment

ANTONY JOHN KUNNAN AND KIRBY GRABOWSKI

KEY QUESTIONS

- What is large-scale second language assessment?
- How are large-scale language assessments developed?
- How do teachers know if an assessment is useful for their purposes?

EXPERIENCE

Sarah is an English as a second language (ESL) teacher in an intensive English program (IEP) in an American university. Many of her students are recent immigrants whose first language is not English. While some of these students are interested in studying in a university after graduation, they have just learned that they must take a large-scale English language proficiency assessment to be considered for admission. They understand the high-stakes nature of the assessment, so Sarah’s students want her help in understanding the purpose of the assessment: who the users of the information might be, what the test assessment is like, and how the results are likely to be interpreted. Sarah, therefore, needs a working knowledge of second language assessment theory and practice before she can address their concerns.

Language teachers-in-training are generally required to take courses in language acquisition, teaching methodology, curriculum and materials design, specific teaching skills (e.g., listening, speaking, reading, and writing), pedagogical grammar, discourse analysis, the use of computers, and so on. However, many teachers do not receive much in the way of courses or training in the field of language assessment, either in terms of theory or practical training. A grounded understanding of the principles of language assessment is crucial for teachers in every context of language instruction.

For example, in lesson plans, teachers create teaching or learning targets (or objectives) that need to be met; assessment then helps them figure out whether (and to what extent) those targets have been met. Without assessment, teaching would be incomplete and it would be nearly impossible for students to have any systematic indication of their abilities with respect to what they have been taught. Traditional classroom assessments (e.g., pre-unit checks, essay assignments, homework, midterms, and final exams) are small-scale assessments that teachers need to be familiar with. In addition, both inside and outside school settings, large-scale assessments are increasingly being employed by stakeholders (e.g., teachers, school administrators, employers, and governments) to gather information about what learners know and can do in a second/foreign language. Further, given the wide array of assessments in use today, there are a number of theoretical considerations that teachers-in-training should understand so they can interpret the usefulness and trustworthiness of the data obtained from their own assessments and from those that they may have to use in their teaching contexts. To this end, in this chapter we describe the major theoretical underpinnings of second language assessment and how these concepts support and inform assessment use. We focus here on large-scale assessments in particular; however, most, if not all, of these concepts can, and should, apply to small-scale assessments as well. Our aim is to help
classroom teachers better understand the various considerations related to language assessment so that they may make more knowledgeable and effective decisions about instruction and assessment.\footnote{1}

**WHAT IS LARGE-SCALE SECOND LANGUAGE ASSESSMENT?**

Language assessments are ubiquitous and can be found anywhere, from elementary schools to high schools, colleges and universities, the workplace, and even immigration and naturalization contexts, and at all levels, from beginner to advanced levels of language teaching and learning. In these different contexts, the language assessments used could be large-scale assessments (such as those prepared by professionals in testing agencies, a university, or a state board or ministry of education) or small-scale assessments (such as those prepared by a teacher, an assessment development group or committee, or a single immigration or naturalization examiner). Teachers in all these contexts need to understand the main characteristics and advantages and disadvantages of large- and small-scale assessments.

**Large-scale versus small-scale assessments**

Large-scale language assessments, also known by the traditional term *standardized tests*, are most often used in school contexts for entrance or exit purposes and to monitor student progress through standardized development, administration, scoring, and reporting. Large-scale assessments are also used to collect uniform baseline information from a large group of students, provide diagnostic information to all stakeholders (e.g., teachers, students, parents, and school administrators), and ensure state-level accountability. An example of this type of assessment is the California Standards Test (CST) (Educational Testing Service, 2011) in English language arts and many other subjects, which is administered to school students from grades 2 to 11. The CST is used to “measure students’ progress toward achieving California’s state-adopted academic content standards in English-language arts . . . which describe what students should know and be able to do in each grade and subject tested” (Educational Testing Service, 2011, para. 1).

Another type of large-scale assessment at the college and university levels is known popularly as the entrance examination. This type of exam is used primarily in the screening and selection of applicants to these institutions, and it typically measures student achievement or language proficiency for accountability purposes, to encourage competition, and to ensure equal opportunities because only the highly able can be rewarded with admission to colleges and universities and job opportunities. In the United States, the Scholastic Achievement Test (SAT) (The College Board, 2012) is an example of an assessment that measures ability in three areas: critical reading, writing, and mathematics. In the area of second language assessment, international examples of large-scale assessments include the Test of English as a Foreign Language: Internet-Based Test (TOEFL iBT) (Educational Testing Service, 2012b); the International English Language Testing System (IELTS) (University of Cambridge ESOL Examinations, 2009–2011); and the Michigan English Language Assessment Battery (MELAB) (Cambridge Michigan Language Assessments, 2012). These three well-known tests measure the English language proficiency of test-takers hoping to study in English-medium universities. A fourth well-known second language assessment, the Test of English for International Communication (TOEIC) (Educational Testing Service, 2012d), is used by employers around the world to measure how well test-takers can comprehend and read workplace English as they listen to and read it. (See the Appendix for descriptions of major large-scale English language proficiency assessments.)

As varied as their purposes are, the main feature of large-scale language assessments is the uniformity (or standardization) of the assessments and testing practices, including development, administration, scoring, reporting, and score interpretation across geographical regions, administration time, and human raters. This has been possible largely because of modern educational measurement (psychometric and statistical) theory and experience with large-scale assessment practices, particularly where test reliability is concerned. More specifically, reliability refers to consistency of measurement, usually across test items or tasks, forms, occasions, and raters. For example, if different test forms can be considered
equivalent, and if raters assign similar scores to test-taker performance, then the test can be said to have higher reliability. However, due to the numerous resources involved in systematizing test development, administration, and scoring, reliability is more easily maximized in large-scale testing contexts. While the emphasis on uniformity has served large-scale language assessments reasonably well, there have also been criticisms regarding the inflexibility of conceptualization and response format (most of these assessments mainly use the selected-response format, e.g., multiple choice) resulting in inadequate skills diagnosis and feedback to test-takers and score users. This affects the validity of score interpretations of an assessment—whether an assessment is meaningful, relevant, and sufficient for the purpose of the assessment. In response to these criticisms, many large-scale test developers have been working to address these inadequacies in recent years by introducing test tasks that require test-takers to produce written or spoken language. While these types of tasks require more resources and may compromise the reliability of the measurement to some extent, the claims about test-taker ability being made from test performance (the validity of score interpretations) are ultimately more defensible.

In contrast, small-scale language assessments are most often used in school, college, and university settings, where the assessment is used to monitor student progress or achievement, typically by the classroom teacher, or in workplace or immigration and naturalization settings by individuals. Examples from the U.S. context are teacher-made classroom assessments for grade-level students (e.g., unit tests, midterms, and final exams), employer’s face-to-face interviews, and immigration examiners’ interviews for the U.S. naturalization test. Although this last example is a high-stakes assessment (i.e., a test that informs life- or career-changing decisions), it is still relatively small in scale, typically conducted in one-on-one and face-to-face settings. An obvious advantage of small-scale assessments is that items can be written or delivered by the teacher, instructor, or examiner face to face so that the assessment is directly relevant in terms of the course content or program. In addition, scoring guidelines can be devised with the assessment’s specific purpose in mind. And, perhaps most important, direct diagnostic feedback, including strengths and weaknesses regarding the performance, can be provided to the test-takers on an individual basis. The main disadvantage of this type of assessment is the variability in assessment design across teachers, instructors, and examiners in terms of items, scoring, reporting, and decision making, which may compromise the reliability of the measurement. This lack of uniformity can often make comparison across assessments and assessment contexts quite difficult. (See Katz, this volume, for an in-depth treatment of small-scale or classroom-based assessment.)

Norm-referenced and criterion-referenced approaches

There are two traditionally held frames of reference, or approaches, to large-scale assessments: norm-referenced and criterion-referenced. Many aspects of assessment (such as development, scoring, reporting, and research) are carried out differently depending on the approach.

To illustrate the difference between norm-referenced testing (NRT) and criterion-referenced testing (CRT) approaches, let us present two different scenarios. First, imagine that a graduate program has 20 scholarships available for newly admitted international students each year. These scholarships are granted to the top 20 students based on their performance on a language proficiency exam. Thus, rather than being determined by a predetermined cut-off score, the scholarships are given to the 20 top-performing students regardless of their numerical score on the assessment. In other words, an absolute level of proficiency (i.e., a certain score) is not required; rather, it is the top 20 test-takers’ performance in relation to all other test-takers that determines their relative standing. In fact, the score of the lowest-achieving scholarship grantee could change from year to year, depending on the relative performance of the top 20 test-takers. In this example, the assessment approach is norm-referenced since the interpretation of assessment performance is relative rather than absolute.

Compare the first example with a second scenario, which involves the New York State English as a Second Language Achievement Test (NYSESLAT) (New York State Education Department, 2011). ESL students in the public school system in New York state must take the NYSESLAT to track how
well they are progressing with English and also to determine whether they are proficient enough to be included in regular English-speaking classrooms. In this case, New York state has a predetermined cut-off score that the students must meet or surpass to be exempt from taking ESL classes. In this case, the cut-off score represents an absolute level of mastery that the students must achieve for them to be considered to be at a high enough level to go directly into regular mainstream classes without additional ESL instruction. In this example, the testing approach is criterion-referenced since the interpretation of test performance is absolute rather than relative to other students’ performances. Similarly, all teacher-made classroom assessments and immigration and citizenship tests are typically CRTs since it is the ultimate level of mastery of the material on the assessment that is of primary concern rather than the test-takers’ relative standing compared to other test-takers. In recent years, many testing agencies have followed suit and adopted a criterion-referenced approach in their large-scale, high-stakes English language proficiency tests. Large-scale test users are gradually coming to an understanding that what they are most interested in is the test-takers’ level of mastery rather than one test-taker’s performance in relation to another’s. Although the type of interpretation of assessment performance (relative vs. absolute) is not the only consideration in determining the distinction between a NRT and a CRT, it is arguably the most salient one in a majority of cases.

CONCEPTUAL UNDERPINNINGS

Assessment development

How is a large-scale language assessment developed? The development of a language assessment, whether it is done by individual teachers or a multinational testing conglomerate, needs careful preparation and expertise in assessment conceptualization, blueprint design, item/task writing, and research. Taking cues from Bachman and Palmer (2010), Fulcher and Davidson (2009), and the example of the TOEFL iBT development (Jaimison, Eignor, Grabe, & Kunnan, 2008), we frame here the cyclical and iterative nature of assessment development in terms of questions and answers so that teachers can understand the purpose of these tests, how they are created, and how they are administered.

We discuss next these relevant questions: What is the purpose of the assessment? Where does assessment development begin? How is assessment format and content planned for in a systematic way? When might technology be used? What does item/task quality control look like? What research is conducted as part of assessment development?

What is the purpose of the assessment? The purpose of large-scale assessment has traditionally been to gather information about test-takers’ ability so that placement, proficiency, certification, or achievement decisions can be made. Increasingly, however, information from large-scale assessments is being used for diagnostic purposes—namely, to assess test-takers’ strengths and weaknesses so that future teaching (or learning) can be prescribed. Once test users clarify the primary purpose of the assessment, the assessment can be organized to fit the purpose. The general scope of the assessment can then be planned in terms of constructs (i.e., the theory that refers to what is being assessed), skills that need to be assessed (i.e., listening, speaking, reading, writing, and/or integrated skills). In addition, information about the test-takers (ability level, age, gender, and any other salient information) is considered. Finally, teachers need to know how an assessment is scored, interpreted, and used for decision making so that the purpose of the assessment is served. These are essential aspects that need to be identified in the planning stage.

Where does assessment development begin? The first step in answering this question is to gather information regarding the target-language characteristics and tasks that are the focus of the assessment, possibly identified in the content standards or the curricular objectives for the course. Most curricula at the school level have these standards or objectives articulated in school-district scope and sequence documents or at the national level by ministries of education. In other situations, the adopted textbook(s) can be used for this purpose. Test developers identify which of the standards, objectives, or textbook materials need to be part of a particular assessment, keeping in mind the purpose of the assessment. For example, they determine if it is important that the test-takers show a mastery of the standards (as opposed to
a sample) to indicate that they have learned the material sufficiently. Another way assessment content is identified is through the characteristics of the target-language use (TLU) domain. Defining the TLU domain involves a specification of how the target language should be used by the test-takers in a particular domain in terms of language tasks and language characteristics. Specifically, the TLU domain can be described in terms of the linguistic features, skills, or functions that the test-takers need to know. Attempts are made at this stage to identify authentic and communicative language items/tasks that relate to the ways in which the test-takers need to use the target language in their real lives. Examples of this include the ability to read for information or to listen to a teacher’s introduction to a topic. These considerations are fundamental to the integrity of the test-design phase.

How are assessment format and content planned for in a systematic way? Once the TLU domain is clearly identified, the stage is set for the development of a general blueprint for the assessment. This blueprint can then be taken forward into the writing of specifications that include specific details regarding the constructs, skills, or abilities in terms of task types, the format of the input (e.g., item stems such as an incomplete statement or direct question, reading passages, or lectures), selected-response formats (e.g., multiple choice, true-false, matching information in columns, gap filling), constructed-response formats (e.g., short responses, such as words or phrases, to extended discourse, such as an essay or a speech), the number of tasks, the overall time allotted for the tasks, delivery matters (e.g., paper-and-pencil, computer, or Internet delivery), accommodations for test-takers with disabilities, and administrative aspects (e.g., security issues). These are also an essential part of the design stage. The more detailed the specifications, the easier it will be for the item/task writer; this is particularly important if multiple writers are involved in item construction and review or if more than one form of a test must be created. Once the blueprint and item/task specifications are ready, they can be handed over to the item/task writing group or committee.

A number of approaches are available to writers when planning to write items/tasks for an assessment. One option is to write items based on the item/task specifications (discussed earlier); a second approach is to write items/tasks based on content standards or curriculum objectives already outlined that may be readily available; and a third, but somewhat more difficult approach, is to develop a theory of the language knowledge required for the ability level that the assessment is being constructed to measure (e.g., listening and speaking for air traffic controllers or writing for business or engineering students) and then to operationalize this into items/tasks. Scoring considerations should also be part of the planning so that the actual items/tasks can be scored in an appropriate manner. Specifically, decisions regarding whether the scoring will be based on dichotomous scoring (correct/incorrect), partial credit, or a rubric have to be made. This is a critical component of the operationalizing stage.

When might technology be used? The answer to this question depends on whether test-takers and teachers have access to computers and Internet connections in the test-taking locations and whether the test-takers are familiar with using a keyboard, mouse, and other technological devices. If this is the case, then assessments can be planned and delivered with the help of computers. In terms of assessment materials and response format, computers can be used to include multimedia materials (e.g., text, photographs, audio, and video) in the test and to require test-takers to perform tasks based on the mouse and keyboard, such as keying in a response, speaking into the microphone, and using mouse clicks, screen touch, and drag and drop. Scoring items/tasks can also be automated for the selected-response as well as for longer response formats.

In spite of the allure of using technology in assessment contexts, researchers have raised questions about whether the use of computers has altered the test-taking process. For example, in terms of writing, one question is whether keying in an extended response (with the other advantages of computers, such as cutting and pasting and using the grammar and spell-check functions) is the same as writing on paper (see Kunnan, 1999). Other issues include impersonation, plagiarism, software compatibility, and computer reliability. Perhaps, these questions will become outdated as more academic and professional work is done with the help of computers and the ease of using the relevant technologies takes care of other concerns. In the meantime, however, these are an additional but critical consideration during the design stage.
What does item/task quality control look like? There is usually an extensive content review process that happens when a new test is being developed or when new items/tasks are introduced into an existing test. During this process, expert reviewers, who know the assessment specifications, content standards, or curriculum objectives, have the opportunity to provide comments and suggestions. Their focus typically includes the appropriateness of the materials, task types, and scoring guidelines; they also check for typographical errors and the layout of the items/tasks. The pre-testing of items/tasks is also useful if sample test-takers (who are representative of the target population) are readily available and it is convenient to administer the assessment to these individuals. Performances during pre-testing can then be used to assess several qualities of the items/tasks, including difficulty (how hard or easy an item is), discrimination (how hard or easy an item is for high- and low-performing test-takers), quality of the response choices (if there are multiple-choice items), time allocations for items/tasks, and ease of scoring the tasks (checking for clearly acceptable answers). Once information on all these aspects is available, each item is usually scrutinized before it is accepted, rejected, or revised. If both expert content review and pre-testing are used, the quality of items/tasks should be acceptable. The accepted items/tasks can then be assembled into a single assessment, keeping the general specifications in mind. This is also an essential part of the operationalizing stage.

What research is conducted as part of assessment development? Although research is often considered last in the cyclical process, it is a fundamental part of the development of an assessment. It is part of the operationalizing stage when pre-testing of items and tasks is done. Typically, the main aim of research during pre-testing is to assess the suitability of items/tasks for inclusion in an assessment. This is done based on item/task difficulty and discrimination, two statistical indices that are used for this purpose. In addition, at this stage, research is often conducted on various other components such as timing for each item/task, scoring points, and weights. Similarly, research on test-taker performance (i.e., of the entire sample or of subgroups) can reveal if there are low performances on certain items/tasks or for the whole assessment. Further, if research does not confirm that the overall consequences are beneficial to the test-takers, then the plan and design may need to be reviewed. Ultimately, findings from such focused research studies may result in adjusting or rewriting the basic plan, blueprint, and/or specifications.

Assessment development as a cyclical process. Although the above six questions might seem quite separate, the development of an assessment is an integrated and cyclical process, starting generally with the planning of the assessment and ending with decision making. Figure 1 illustrates the main stages or activities involved in the development of an assessment: planning, designing, operationalizing, using, and researching. It also shows the multidirectionality of the stages in a real assessment development and use cycle.

During the planning stage, the purpose of the assessment is established along with the intended consequences of test use, including the potential decisions that can be made and the impact on test constituents. The designing stage includes identifying the TLU domain and outlining the test blueprint and specifications, including any technology considerations. Item- and task-writing, pre-testing, and revision then follow in the operationalizing stage. Once the test is

![Figure 1. The cyclical nature of assessment development.](image-url)
ready, it can be administered and scored as part of the using stage. During this stage, score-based inferences can be made, which translate into decisions about individuals or groups. Once test developers get data from a test administration, research on the items or tasks can be performed during the researching stage. Findings from this research can then inform the planning stage in terms of showing support for or evidence against the assessment’s intended purpose or the perceived consequences of test use. Research findings can also inform changes to the designing stage for future administrations of the test. Note that, since this process is not linear but, rather, cyclical and iterative in nature, each ensuing stage in the process can also inform prior stages. For instance, pre-testing during the operationalizing stage may uncover issues resulting from the designing stage that need to be addressed before the test can be administered again. Similarly, difficulties revealed during test administration in the using stage may require test developers to modify certain tasks in the operationalizing stage so that listening input, for example, is delivered in a more uniform way.

How do teachers know if an assessment is useful?

If teachers are responsible for choosing or administering a large-scale assessment or if they need to interpret its results in a meaningful way and make decisions based on the scores, they need to understand the fundamental considerations in language assessment research and how this information can be interpreted. Since the selection and evaluation of an assessment should be based on evidence that is available in technical manuals, research reports, or other materials, a firm understanding of test reliability, validity, and fairness is critical. These issues have long been considered foundational evaluative aspects in language assessment research. With some training in understanding language assessment research, teachers should be able to select an appropriate assessment and/or evaluate an assessment being used. To present these concepts, we use here a modified version of the theoretical interpretive and validity arguments (Kane, 1992) and assessment use arguments (Bachman & Palmer, 2010) by presenting the notions of claims and evidence.

Claims and evidence. Claims are assertions or statements that assessment developers (e.g., teachers, committees, or agencies) make regarding an assessment. These claims are generally stated by testing agencies, but they also serve as a good model for teachers who develop assessments. Evidence is defined as any research finding based on the analysis of assessments, assessment performance, and the impact of assessments.

As an example, the TOEFL iBT makes an explicit claim to prospective test-takers about the assessment on its website: “The TOEFL iBT assessment measures your ability to use and understand English at the university level. And it evaluates how well you combine your listening, reading, speaking, and writing skills to perform academic tasks” (Educational Testing Service, 2012c, para. 2). On the other hand, the developer of the California High School Exit Examination (CAHSEE) (California Department of Education, 2011) provides a less explicit claim on its website: “The purpose of the CAHSEE is to improve student achievement in high school and to help ensure that students who graduate from high school can demonstrate grade-level competency in reading, writing, and mathematics” (para. 1).

Regardless of how clearly their claims are stated, assessment developers need to assemble research evidence that supports such claims so that they can justify the use of the assessment. Such evidence could be findings from research that support the many aspects of an assessment. Table 1 presents some example claims, along with acceptable forms of evidence, that an assessment developer might make about an assessment. Once assessment developers have articulated the claims, they need to provide evidence for each of them (or other parties could provide evidence for counterclaims). An example of how this might be done is presented next.

Claim 1: The assessment has beneficial consequences. Kunnen (2004, 2008) notes that all assessments should be beneficial to society and that, in particular, assessments should not be harmful or detrimental. This is especially critical with respect to large-scale assessments, which are often high-stake tests. Arguably the test-takers—the most directly affected stakeholder group—are affected in terms of their preparation (e.g., the change in their cognitive abilities as a result of formal classroom instruction or self-learning), the assessment they
Table 1. Possible Claims and Evidence

| Claims                                                                 | Evidence                                                                                                                                 |
|------------------------------------------------------------------------|
| 1. The assessment has beneficial consequences.                         | Studies of consequences on the test taker, instructional system, and the wider community                                             |
| 2. The assessment is consistent.                                       | Internal-consistency and external-reliability scores                                                                               |
|                                                                        | Consistency across different items/tasks, forms, occasions, and raters                                                            |
| 3. The assessment interpretations are meaningful.                      | Content, criterion, and construct validity                                                                                         |
| 4. The assessment is free from bias.                                   | Research showing assessment free from dialect, content, or topic bias                                                             |
|                                                                        | Research showing interpretations not biased against any test-taker group                                                         |
| 5. The assessment promotes equitable decisions.                        | Decisions based on appropriate standard-setting procedures                                                                       |
|                                                                        | Decisions relevant to societal values and legal requirements                                                                       |
|                                                                        | Decisions equitable to all test-taker groups                                                                                       |

take (e.g., in terms of their affective factors), the scores and other feedback (numerical and/or descriptive) they receive, the decisions that are made about their performance (e.g., grades, pass-fail, certification, permission to immigrate, or granting of citizenship), and the life- or career-changing situations they may experience as a result of the assessment decision (e.g., exemption from an ESL requirement or permission to immigrate). In the case of high school students taking the CAHSEE in grades 9 to 12, test-takers have to pass the mathematics and English language components of the assessment to secure a diploma to graduate from high school and go on to college/university or enter the workforce. Therefore, the consequences of a pass or fail grade on the assessment can have a serious impact on the test-takers. This situation makes it crucial that CAHSEE’s claims and the supporting evidence for the claims be examined carefully so that the decisions that are made regarding test-takers are justifiable.

Claim 2: The assessment is consistent. The meaningfulness of an assessment’s score interpretations (Claim 3) cannot even begin to be investigated without first confirming the consistency, or reliability, of the scores obtained from the assessments. In other words, reliability is a necessary precondition for validity; hence, it is vitally important in determining the trustworthiness of the information obtained from test scores.

As previously mentioned, test reliability can be defined as consistency of measurement. If assessment scores are inconsistent (along one or more dimensions), then the trustworthiness of the results, and therefore the inferences that can be made from them, are necessarily called into question. Ideally, assessment scores are representative of the test-taker’s true ability as defined in the construct. In other words, scores should be construct-relevant and have as little construct irrelevance as possible. For example, a reading assessment should assess material in the texts and not the background knowledge of the test-taker; background knowledge is construct-irrelevant. However, in assessment contexts, there is always a certain amount of error (i.e., construct-irrelevant variance) that is part of a score. There are a number of possible sources of error that detract from the stability of the assessment. These sources depend on the format and content of the assessment and on the context of the testing situation, including assessment method facets (e.g., raters and occasions). Error can be introduced by factors outside the assessment (e.g., human raters), within the assessment itself (e.g., the items/tasks), or by the test-takers themselves (e.g., a test-taker is ill or has test fatigue). The goal is to minimize the amount of error that is a part of the scores so that the assessment scores represent as accurate a measure of a test-taker’s true ability as possible.

Internal-consistency reliability, which is particularly important in justifying the stability of the measurement, relates to the homogeneity of the items/tasks. In other words, internal-consistency
reliability asks this question: Do all the items/tasks in the assessment measure the same thing (e.g., grammatical knowledge or speaking ability)? Internal-consistency reliability, though highly related to validity, does not relate to what the items/tasks are measuring; rather, it is primarily concerned with how much of the observed score variance (i.e., the assessment score) can be attributed to the test-takers’ true score variance (i.e., the test-taker’s theoretical score) rather than error. For example, imagine a 75-question multiple-choice grammar assessment in which there are several items that measure background knowledge rather than grammatical knowledge. These items can compromise the reliability of the assessment since they are introducing construct-irrelevant variance into the scores. In other words, some test-takers who perform very poorly on the grammar items in the assessment may do very well on the five items that have pop-culture references, for which they can draw on their background knowledge (and not grammatical knowledge) to answer these questions correctly. These types of items create anomalies, or inconsistencies, in the test-takers’ responses that are unrelated to their grammatical knowledge. This could potentially compromise the reliability of the assessment as a measure of grammatical knowledge. This type of unwanted variability, though it can never be completely eliminated, can be minimized through systematized assessment design and development procedures.

External reliability relates to the factors outside the assessment itself (e.g., different forms or occasions) that have a potential effect on the consistency of the scores. For instance, in large-scale testing, test-takers ideally would receive the same score on an assessment irrespective of which version (Test Form A or B) they took or on which occasion (in January or in June) they took the assessment. If they do, there is evidence that the assessment forms or occasions can be considered interchangeable since these variables introduce little error into the measurement.

Another potential factor affecting external reliability is the lack of consistency among the human raters of extended speaking and writing tasks. Although there is a benefit to performance-based tasks in communicative language teaching contexts, in that test-takers can be engaged in communicative activities similar to those used in the classroom (such as interviews, conversations, and extended writing), human raters are required to score the response samples. These raters inevitably introduce variability into the scores, which can compromise reliability. Rater consistency can be maximized through highly specified domain descriptors in a scoring rubric, extensive rater training/norming, and a deep understanding of the construct being measured. If raters exhibit a high level of consistency, there is evidence that the scores can be interpreted as being a trustworthy indication of the test-takers’ true ability. (See Katz, this volume, for examples of analytic and holistic rubrics used for scoring writing assignments.)

Claim 3: The assessment interpretations are meaningful. Once test reliability has been established, arguably the most important quality of an assessment is the extent to which the score interpretations are meaningful. This is commonly known as validity, and it is at the core of assessment research. Evidence for validity can be gathered from a variety of sources, but it is not necessary to use all of them. Nevertheless, the more evidence that can be provided, the stronger the argument for the meaningfulness of the interpretations. The three predominant types of evidence that are collected are content-related evidence, criterion-related evidence, and construct-related evidence, though all are often inextricably intertwined.

Content-related evidence of validity, the first type of evidence that is collected to support this claim, is determined based on an assessment’s item representativeness (i.e., Does each individual item belong on the assessment?) and domain coverage (i.e., Is a sufficient amount of the content we are trying to measure represented in the assessment?). For example, if the purpose of an assessment is to measure students’ mastery of second language reading skills taught at the sixth-grade level, to maximize item representativeness ideally all the assessment items should contribute to making inferences about the test-takers’ reading ability at a sixth-grade level. Items that do not measure knowledge of reading ability at a sixth-grade level do not show good item representativeness and therefore compromise the meaningfulness of the assessment score interpretations. With respect to domain coverage, if the purpose of an assessment is to measure knowledge of the linguistic structures covered in Units 1–3 of a given textbook, we want to make sure that a sufficient amount of the Unit 1–3 content is represented in the assessment. Having
too little content coverage compromises our ability to make valid inferences about the test-takers' knowledge of those structures. Content experts, including teachers who are familiar with course content or standards, can help determine the adequacy of item representativeness and domain coverage in an assessment.

Criterion-related evidence of validity is collected to support this claim in terms of the capacity of an assessment to predict future performance (e.g., in college or the workplace). For example, if an assessment is used to select high school students for admission into a college or university (such as the SAT in the United States), the expectation is that the SAT is able to predict to some extent the test-takers' readiness for and subsequent performance in college or university courses. If the prediction rate is high, we can state that there is evidence for criterion validity; if the prediction rate is low, we have little evidence for it.

Construct-related evidence of validity is the most important of all types of evidence. Construct validity refers to the extent to which the score-based interpretations are meaningful in terms of the blueprint and specifications, content-standards, curriculum objectives, or theory of knowledge that the test purports to measure. Therefore, a clear and well-defined construct of the ability being measured should be at the heart of every language assessment. Construct definitions can be based solely on a theory of language, or they can be defined in terms of a series of standards or objectives, a syllabus, and/or textbook (all of which are also ideally based on a theory of language). When inferences about test-takers' ability are made based on test scores, evidence is needed that these interpretations are meaningful and appropriate to a given assessment context. Perhaps most notably, construct validity relates to the extent to which the test-taker performance can be generalized to the particular TLU domain that goes beyond the relatively limited scope of the assessment itself. Thus, construct validity involves the ongoing process of justifying the interpretations that we make about test-takers’ knowledge in some real-world context based on their scores.

Claim 4: The assessment is free from bias. The topic of unbiased interpretations has been shown in the last few decades to be an essential part of administering a fair assessment. The Standards for Educational and Psychological Tests and Testing (from the American Psychological Association, American Educational Research Association, & National Council on Measurement in Education, 1999) first articulated fairness in testing and test use. In his Test Fairness Framework, Kunnan (2004) further delineates test fairness in terms of the types of bias assessment developers should focus on: (1) dialect, content, and topic; and (2) group performance. Since potential systematic bias in the dialect, content, or topic can exist in an assessment, an examination of the assessment is necessary to determine if bias exists. When examining each item, we ask this question: Do any of the items/tasks or texts have the potential for systematic bias (given the test-takers)? For example, imagine an English language assessment being used in the United States to determine whether English language learners can be mainstreamed out of their ESL classes in an elementary school. In this assessment, learners listen to texts with conversations from speakers who speak different dialects of English (American English, British English, Australian English, etc.). This is a source of dialect bias if the learners have not previously been exposed to these dialects and they are now being assessed for their ability to understand the conversations. Similarly, imagine an assessment that requires test-takers to read material or write about content or topics that they are unfamiliar with. For example, imagine that students are asked to write an essay about how they enjoyed the Chinese Lunar New Year or Cinco de Mayo celebrations. This kind of task is a source of content or topic bias if specific cultural background knowledge is expected and the test-takers do not have it (i.e., are not Chinese or Latino or have never experienced such a celebration). Similar concerns can occur in reading assessments where obscure words that are known by some and not by others (e.g., from pop culture, rocket science, or jazz) are included in the assessment. Content analysis of items/tasks in all sections of an assessment should be conducted before the assessment is administered to ensure the absence of bias. If this is not done, the test-takers will have been unfairly penalized and their scores may not be a meaningful representation of their abilities.

It is also possible that different subgroups of test-takers (e.g., in terms of gender, age, race/ethnicity, or native language) may perform differently on an assessment. What we want to know is whether the difference in performance is based on construct-relevant factors (e.g., grammatical
knowledge or writing ability) or irrelevant factors (e.g., gender or socioeconomic status). For example, to examine a high school writing assessment to uncover any gender bias, we would check to see whether the male test-takers have a similar profile of scores as the female test-takers. In some instances, we may find that they do not. For example, a particular item or task that has traditionally male- or female-oriented topical content may turn out to be very easy for one gender group but not for the other. For example, imagine a scenario in which male test-takers are interested in football and female test-takers are interested in golf. In this situation, test content in a reading assessment with information on football would likely favor male test-takers and test content with information on golf would likely favor female test-takers. When such differences occur systematically and are statistically identified as such, these texts and items/tasks are flagged and subject to a content review to ascertain the source of bias, if any. Results from such investigations could result in the review, modification, or deletion of items/tasks if they are found to have bias.3

Claim 5: The assessment promotes equitable decisions. Examples of decisions that are made about test-takers based on test scores include whether they succeed in passing a course or program, earning a diploma, being offered employment, securing permission to immigrate, or being granted citizenship. In school, college, or university contexts, setting standards for different levels of performance (e.g., A-F grades or other marking systems) is also important because test-takers need to be awarded appropriate grades based on their performance and instructors/teachers need to know whether they have created classification errors (have given a test-taker a higher or lower grade or placement than was appropriate). Decisions based on test-score interpretations are expected to be equitable based on appropriate standard-setting procedures (e.g., defining a level of achievement or proficiency that corresponds to a certain standard-setting score or cut score), societal values, and also the legal requirements of the community in which an assessment is administered. Again, the higher the stakes of the assessment, the more important equitable decisions become.

Another concern in decision making is whether the decisions that are made are equitable to all test-taker groups. This is particularly important if standard setting and subsequent decision making are not empirically defensible but are carried out following public policy or practice. For example, in a community where meritocracy is the preferred decision-making process, test-takers with the highest scores are the ones who receive the benefits. On the other hand, in a community that has a policy to assist individuals who have been discriminated against in the past, test-takers without the highest scores may be the ones to receive benefits. In many countries, such a policy has resulted in individual standards for different test-taker groups based on race/ethnicity, gender, or religious membership. Any decisions affected by legal requirements (e.g., court rulings, settlements, and precedents) should match the societal values of the community, but all decision making, whether it is merit-based or quota-based, has to stay within the letter and meaning of the law.

Counterclaims. Counterclaims can be articulated by interested parties such as test-takers, schools, parents, and school districts. Some possible counterclaims include that the assessment does not have beneficial consequences, the assessment is not consistent, the assessment interpretations are not meaningful, the assessment is not free of bias, and the assessment does not promote equitable decisions. Evidence that supports each of the counterclaims will then have to be produced in order to arrive at a public or judicial resolution.

CLASSROOM APPLICATIONS

Now let us think back to our IEP teacher Sarah, described in the Experience section of this chapter. Recall that Sarah’s students are recent immigrants to the United States and must take a large-scale English proficiency test to be considered for university admission. Now that Sarah has a working knowledge of second language assessment theory and practice, she can better understand the test and help address her students’ concerns. We next pose questions that are relevant for Sarah to answer.

What is the purpose and content of the assessment?

The purpose of the high-stakes assessment is to measure non-native test-takers’ language proficiency. Language proficiency on the test is defined in
terms of how English is heard, spoken, read, and written in the university classroom. The assessment is a large-scale, high-stakes, criterion-referenced test.

**How can Sarah interpret the information using what she has learned?** Since the assessment is a criterion-referenced test, what is most important is not how well the test-takers do in relation to other test-takers; rather, it is solely the test-takers’ proficiency level that determines their score. This means that Sarah’s students must ultimately improve their English to do well on the test.

**Who are the users of the information that will been gathered from the assessment?**

Admissions officers and/or academic departments or programs often require non-native speakers of English to show evidence of their ability to perform at a high enough level to study in the university environment. Thus, they set cut-off scores for these potential non-native-speaking applicants based on the demands of the curriculum to either accept the applicant for admission or to indicate that the applicant needs additional language training.

**How can Sarah interpret the information using what she has learned?** Cut-off scores are usually set for the entire test, but minimum scores may be required for individual skills sections as well. If test-takers know that they are lacking in one ability or another, it is important for them not only to prepare for the test so that they can surpass the cut-off score but also to improve their language ability so that they can ultimately become a successful student if admitted to a rigorous academic program.

**What is the content of the assessment?**

The assessment has four sections: reading, listening, speaking, and writing. The speaking and writing sections have both independent and integrated tasks. For the reading section, test-takers will read three or four passages from academic texts and answer multiple-choice questions. For the listening section, test-takers will listen to lectures, classroom discussions, and conversations and answer multiple-choice questions. The multiple-choice questions will assess the students’ ability in various reading and listening skills. The writing and speaking sections of the assessment have two different types of tasks, independent and integrated. In addition to one independent-writing and two independent-speaking tasks where test-takers write or speak in response to a prompt, there are four integrated-speaking tasks and one integrated-writing task that require test-takers to read and/or listen before responding to a question.

**How can Sarah interpret the information using what she has learned?** Since the purpose of the assessment is to measure academic language proficiency, the test tasks are meant to represent how language is used in the TLU domain—the university context. From the test specifications, it seems that the reading and listening sections, and the independent-speaking and -writing tasks, are fairly straightforward and conventional, and will likely be familiar to the students. Perhaps somewhat unfamiliar, though, are the integrated tasks, which are meant to capture the way academic language skills are combined in an integrated way in the university context. For example, students often read a text for homework, come to class and listen to a lecture about that same topic, and have to summarize that information for a friend who missed class. Or a student might listen to a conversation between two friends about a controversial new university policy and then join the discussion and argue her own point of view. So, in the actual assessment, a test-taker might be asked to read a short passage on an academic subject, listen to a lecture on the same topic, and then respond orally to a question about what he or she has read and heard. Or a test-taker may be asked to listen to a conversation about a problem, summarize that conversation, and then offer his or her solution. These are the types of tasks Sarah’s students can expect and should prepare for.

**How can users be sure that assessment results are indicative of the test-takers’ abilities?**

Test-score users want to be sure that the scores obtained are a precise and trustworthy representation of the test-takers’ ability. Thus, they are relying on the assumption that the test developers have evidence for the claims they are making about the test, including reliability, validity, and fairness considerations.
The test developer will have a full research program dedicated to making sure that the assessment’s scores are precise and that there is sufficient evidence supporting the valid interpretation and use of the test scores. First, the internal-consistency reliability of the assessment could be above 0.90, which is very high. Second, the validity of the assessment could have been a major consideration from the conceptualization and design phases of the test, including the relevance and representativeness of test content, task design and scoring rubrics, the construct definition underlying the test, and the consequences of test use. Last, the test developer routinely investigates any bias present in the test.

How can Sarah interpret the information using what she has learned? The high reliability is, in part, a result of systematic test development procedures, particularly where test specifications are concerned. The test developer also has an extensive and rigorous rater-training program where raters are continually normed. This also undoubtedly contributes to the high precision of test scores. To ensure that the test content is relevant and representative, the test design process includes the analysis of the academic tasks needed for study at English-medium institutions of higher education and the identification of important characteristics of these tasks that can be captured in the test. With respect to test development and scoring design, task types, task characteristics, and the user interface are all important considerations. Perhaps most important, research on the assessment shows evidence that the test measures a complex and multicomponental (i.e., reading, listening, speaking, and writing) construct of ESL/EFL ability, consistent with what the test was designed to measure. In terms of the positive consequences of test use (i.e., positive washback), in preparing to take the test, test-takers may be exposed to the use of test preparation materials and activities that more closely resemble communicatively oriented pedagogy in academic English courses, thus improving their language skills. From a global perspective, the assessment also affords non-native English speakers the opportunity to demonstrate their language ability when applying to English-medium universities.

Thus, Sarah will be better prepared to plan her materials and instruction in her IEP program by keeping these points in mind. She can also make these ideas part of a mini-test preparation course she might plan in addition to her regular course, which could include instruction on how best to approach the test tasks and also practice tasks that are similar to those on the test itself.

FUTURE TRENDS

Computer technology

With the wide availability of computer technology, assessment practice is employing computers and computer technology in various ways. One simple way is to use computers to deliver assessments. In this approach, known as computer-based tests or computer-delivered testing, the computer merely replaces the paper-and-pencil assessment for displaying the test materials and collecting responses. Another approach that is more complex is to use computers to assemble assessments for different test-takers in a tailored fashion; this is known as computer-adaptive testing (CAT). In this approach, items or tasks are matched in difficulty to a test-taker’s ability based on the test-taker’s previous correct or incorrect responses. Although there are at the time of this writing very few well-known language assessments that use CAT—the Oxford Online Placement Test (OOPT) is one (Oxford University Press, 2011)—this approach is becoming popular in educational assessment. Both these approaches assume that test-takers have access to computers to take the tests.

Another use of computers is in the computerized/automated scoring of constructed-response tasks in both writing and speaking assessments. In writing assessment, computers are trained to score short responses that require a word, phrase, or sentence. Computers are also being programmed to score longer responses such as essays as a way of possibly supplementing or ultimately replacing human raters with computer technology; for example, e-rater® for the TOEFL iBT independent-writing task is currently in use (Educational Testing Service, 2012a). Similarly, in speaking assessments, computer technology is being used to score short spoken responses (see Versant tests for details; Knowledge Technologies, 2011).

A related matter is the issue of whether human readers rate keyed-in or typed versions
of essays more favorably than handwritten ones. Research evidence from experimental studies has found that handwritten essays received significantly higher scores than their typed counterparts (Powers & Farnum, 1997; Russell & Tao, 2004). Why human readers rate handwritten versions of essays more favorably than the typed versions is unclear, but perhaps rater expectations may be higher with the typed versions.

In brief, we can expect an increased use of computer technology in language assessments in the next decade. This will also bring new challenges to the profession because we need to be sure that the benefits from technology do not negatively impact assessments and test-taker performance.

New attempts at diagnostic feedback
A recent trend among psychometricians and test developers is to provide diagnostic feedback and profile reports to test-takers. Three large-scale language assessments, TOEFL iBT, IELTS, and MELAB, provide score reports, and surely more tests will follow suit. These reports include individualized test-taker feedback, which is a vast improvement on the basic set of scores that test-takers were provided with in the past. Kunnan and Jang (2008) provide some discussion of the main issues related to diagnostic feedback from large-scale assessments.

Another assessment tool whose primary aim is in diagnosing rather than selecting or certifying language proficiency is DIALANG (Lancaster University, 2006–2012), which is a computer-delivered diagnostic language assessment system. In this system, test-takers can take tests in reading, listening, writing, vocabulary, and grammar in 14 European languages. Test results are reported against the six levels of descriptors of communicative activities in the Common European Framework of Reference for languages (Council of Europe, 2001).

Skills integration
Although most language assessment tasks and response formats have remained the same over many decades, a few innovations have been implemented recently. For example, the TOEFL iBT has integrated-skills tasks (read, listen, and then speak in response to a question; and read, listen, and then write in response to a question). This is a significant change because most assessments have continued to assess language abilities in terms of independent skills (listening, speaking, reading, or writing) despite calls for more authentic assessments that include skills integration.

CONCLUSION
In this chapter, we have presented an introduction to the field of second language assessment with a focus on large-scale assessment. Although the current chapter merely scratches the surface of the field, it is our hope that, at the very least, it provides a foundation in terms of assessment practice and research. A further discussion of test design and development considerations, including test specification and item-writing guidelines, is necessary to understand assessment development more fully. Furthermore, from the research side, both quantitative data-analysis procedures and qualitative data-analysis procedures are certainly necessary for a more complete education in language assessment. Nonetheless, we feel that, given the information in this chapter, teachers-in-training should feel better equipped to make judicious decisions about what assessment practices they choose to use in their own classrooms and to be informed consumers of large-scale language assessments.

SUMMARY
- The distinguishing feature of large-scale language assessments is the uniformity (or standardization) of the assessments and testing practice, including development, administration, scoring, reporting, and score interpretation across geographical regions, administration time, and human raters.
- Norm-referenced and criterion-referenced testing approaches are different in the way that test performance is interpreted—relative in norm-referenced testing and absolute in criterion-referenced testing.
- Assessment development is cyclical and iterative in nature, and includes planning, designing, operationalizing, using, and researching stages.
- When assessment developers make claims about an assessment, they also need to
assemble research evidence that supports their claims, including (at minimum) reliability, validity, and fairness considerations, so that they can justify the use of the assessment.

- Computer technology will undoubtedly influence large-scale assessments in the decades to follow with computer delivery of assessments and automated scoring of constructed-response tasks in writing and speaking.
- Diagnostic feedback and skills integration in large-scale language assessments will make a mark in the next decade.

**DISCUSSION QUESTIONS**

1. What has been your experience with large-scale assessment, either as a test-taker, teacher, or administrator? Did you feel the assessment was fair? Why or why not?
2. Which of the steps in the ESL/EFL assessment development process displayed in Figure 1 might be easy and which difficult to carry out? State your reasons.
3. Define reliability. What is the difference between internal and external reliability? In what types of tests do the different types of reliability described apply?
4. Why is it acceptable for a construct definition to be based on a theory of language or based on a series of standards, a syllabus, or textbook when these all seem like very different things?
5. Have you noticed any texts, tasks, or items on tests that caused content, topic, or dialect difficulties for you? Do you think the assessments should or should not have had those materials? Could you claim the assessment was not fair to some students? Would you have sufficient and relevant evidence to justify this claim?

**SUGGESTED ACTIVITIES**

1. In groups, identify three or four large-scale and small-scale ESL/EFL assessments that you are familiar with. Describe the main characteristics of each.
2. Examine the large-scale and small-scale assessments you described in Activity 1. Identify which of them are norm-referenced and which are criterion-referenced. What characteristics of the assessments make them so?
3. Use a search engine (such as Google or Bing) to look up two ESL or EFL language assessments in an area that you are interested in (such as school or college, workplace, immigration, citizenship). Learn more about the many aspects of the assessment, including development, scoring, reporting, and administration matters. Do you feel that all the information you wanted was available on the assessments’ websites? If not, what additional information would you like to see made public on their websites?
4. Imagine you and a colleague are contacted by your principal or language program director to develop a language assessment in the areas of listening and speaking. Also imagine that you have defined your construct and you have written the test specifications. How would you then decide on the listening and speaking materials (input) for your tasks? Search the web for examples of listening or speaking rubrics. How could you adapt them for use with your own students?
5. Choose a language assessment that you are familiar with, and write one argument that the assessment developers might make for each of the following five claims: beneficial consequences, consistent assessment, meaningful interpretations, lack of bias, and equitable decisions. Now consider what evidence you might need for each of the claims. Is it easier to collect some kinds of evidence than others? Why/why not?

**FURTHER READING**


This book offers the theoretical view that language assessments should be evaluated through a claims and evidence approach.


This book presents in one place the criterion-referenced approach to language assessment (CRT). It presents the theoretical, statistical, and practical aspects of developing, analyzing, and reporting CRT assessments.

This article presents the concepts of test specifications and test development in an innovative way.


This article presents information on the statistical analyses that need to be conducted when analyzing test performance data.

**APPENDIX: OVERVIEW OF SOME MAJOR LARGE-SCALE ENGLISH LANGUAGE PROFICIENCY ASSESSMENTS**

**First Certificate in English (FCE)**

The FCE is one of Cambridge ESOL’s tests, and it measures listening, speaking, reading, writing, and grammar. The purpose of the FCE is to certify that test-takers are able to study, work, or live independently at the intermediate level of English proficiency (B2 in the Common European Framework of Reference [CEFR]). For more information, go to [http://www.cambridgeesol.org/exams/fce/index.html](http://www.cambridgeesol.org/exams/fce/index.html).

**International English Language Testing System (IELTS)**

IELTS is one of the Cambridge English for speakers of other languages (ESOL) examinations. There are two versions of the test: general training and academic. Both test modules measure listening, speaking, reading, and writing. IELTS scores are typically used for university admissions. For more information, go to [http://www.ielts.org](http://www.ielts.org).

**Michigan English Language Assessment Battery (MELAB)**

MELAB, one of the Cambridge Michigan assessments, measures advanced-level English proficiency for individuals applying to educational institutions or for work/training purposes. MELAB measures listening, speaking, reading, and writing. For more information, go to [http://www.cambridgemichigan.org/melab](http://www.cambridgemichigan.org/melab).

**Pearson Test of English (PTE)**

PTE, developed by Pearson Education, is a computer-delivered, skills-based language test measuring listening, speaking, reading, and writing. PTE has three versions: academic, general, and young learners. Each version of PTE is used for different purposes. For more information, go to [http://www.pearsonpte.com](http://www.pearsonpte.com).

**Test of English as a Foreign Language Internet-Based Test (TOEFL iBT)**

TOEFL iBT, developed by Educational Testing Service (ETS), Princeton, is a widely used exam that is administered via the Internet. TOEFL iBT measures academic listening, speaking, reading, and writing through independent- and integrated-skills tasks and is typically used for university admissions. For more information, go to [http://www.ets.org/toefl](http://www.ets.org/toefl).

**Test of English for International Communication (TOEIC)**

TOEIC, developed by Educational Testing Service (ETS), measures test-takers’ English proficiency in the workplace. The test has two versions: one with listening and reading, and the other with speaking and writing. TOEIC is used by employers to determine who can communicate effectively with co-workers and clients when English is the lingua franca. For more information, go to [http://www.ets.org/toeic](http://www.ets.org/toeic).

**ENDNOTES**

1. Community college and adult school ESL teachers also need to understand language assessment theory and practice to help their students in citizenship classes take high-stakes assessments such as the U.S. naturalization test, which has an English language component (see Kunnan, 2009a, 2009b).

2. Evidence of construct validity is very often presented statistically, including estimates of assessment score reliability, variance component estimates, dimensionality, and data-model fit statistics. However, content-related evidence and criterion-related evidence can both serve to help justify and triangulate statistically oriented construct-related evidence because the three are often intertwined.

3. These analyses are generally called differential item functioning and are statistical in nature (see Ferne & Rupp, 2007; Geranpayeh & Kunnan, 2007).

4. Quantitative data-analysis procedures typically entail descriptive statistics, inferential statistics, item-response theory, generalizability theory, and structural equation modeling, while qualitative data-analysis procedures entail item-content analysis, discourse analysis, and verbal protocol analysis.