45

Test Development Literacy

Kirby C. Grabowski
Teachers College, Columbia University, USA

Jee Wha Dakin
Oxford University Press, England

Introduction

In most language-learning contexts, understanding how much learners know or can do in a language is paramount in order to maximize their learning opportunities and make fair and equitable decisions. In the language classroom, teachers are usually the ones who are responsible for making teaching and learning decisions based on assessments. In the context of program placement and large-scale assessment, it is often program administrators, admissions personnel, or even government officials who are making placement, competency, and selection decisions based on test performance. Notwithstanding a strong background in language instruction, program administration, or management experience, these individuals often have little training in test development. If particular stakeholders (e.g., teachers) are actually the ones designing assessments, training in test development is obviously crucial. If stakeholders (e.g., admissions personnel) are using pre-existing assessments, not only is training in test development empowering, but also the knowledge gained will allow them to better understand the nature of the assessment, its reliability, and the validity of any inferences and decisions made from it. Therefore, the purpose of this chapter is to orient stakeholders to various concepts they may need to consider when developing a language assessment. Training in test development should minimally include an overview of the conventional uses of tests, followed by instruction in construct definition and, ultimately, training in test construction, test administration, scoring considerations, and data analysis. Although any of these concepts can be presented singly, given the inter-relationships among them, they are nearly always best understood and made more meaningful when presented together. Last, though the concepts outlined in this chapter will be presented as though they are linear, it is important for stakeholders to understand that test development is a

The Companion to Language Assessment, First Edition. Edited by Antony John Kunnan.

DOI: 10.1002/9781118411360.wbcla129

process that is cyclical and iterative in nature. In other words, test developers often move back and forth between the stages as they gather more information, refine their constructs, revise their test specifications and tasks, and adjust their scoring methods to best suit the needs of their context.

Conventional Uses of Tests

The first thing that any test developer needs to determine is the way in which a test will be used; in other words, what kinds of test-score inferences and score-based decisions will be based on the information gathered from the test. For instance, test users may be interested in stratifying students into ability levels for an adult English as a second language (ESL) program and are, therefore, interested in making placement decisions. A teacher in a language program may be interested in measuring students' mastery over material taught in the class in order to determine whether or not an individual student can pass onto the next level. In this case, achievement or progress decisions are relevant. A university administrator may be interested in making decisions about an international applicant's ability to perform at a high enough level to succeed in a rigorous English-medium academic environment. In this case, competency decisions are relevant. Or perhaps a curriculum developer tasked with creating targeted training materials is interested in diagnostic decisions based on information gathered from a measurement of the strengths and weaknesses of remedial students with respect to their writing ability. These different uses and many others can be categorized into conventional types of tests.

Placement decisions are typically based on information gathered from a placement test. Placement tests should be designed with a particular course of instruction or curriculum in mind (e.g., a conversational English course). In other words, the content on the test should directly reflect the type of course content found within the program in which the test takers will be placed (e.g., conversational English and not academic English). The test content should also correspond to the range of ability of the students in the program itself (e.g., from beginner to advanced levels). This correspondence can help maximize the test scores' alignment with particular ability levels within the program.

Achievement decisions are typically made in instructional domains where the stakeholders are interested in gathering information about the extent of the test takers' mastery over the material taught, or the learners' progress, or both. Traditional classroom tests (e.g., unit tests, midterms, final exams) are all classic examples of achievement tests. They usually measure what students have learned as a result of a certain period of instruction. The test content for these types of tests is generally a direct and fairly narrow reflection of the course material (e.g., textbook, syllabus, teaching or learning objectives).

Selection and gatekeeping decisions are typically based on (large-scale) proficiency tests. Admissions officers or human resources personnel may be interested in gathering information about potential applicants' level of language proficiency to determine whether they are suited to the demands of the coursework or job requirements. The sampling of content for proficiency tests such as these is usually

very broad, is context independent (i.e., the content is not tied to any particular course of instruction), and can be general, academic, or work-related in focus. Scores from proficiency tests are also sometimes used for making program placement decisions or exit decisions if a close correspondence can be shown between the test content and the content from the learning context.

Diagnostic decisions are typically based on information gathered from assessments expressly designed to reveal the test takers' strengths and weaknesses. In terms of test content, test developers usually need to cast a fairly wide net since they may not have specific expectations about what the test takers should know or can do before taking the test. The inferences based on the information gathered can result in decisions about the format, content, or both of teaching and learning on a relatively small scale, such as in a classroom, or they may result in program or administrative reform on a larger scale, such as in transforming a curriculum to meet changing educational standards. Many different types of tests can be used for diagnostic decisions; however, if they are not designed with this purpose in mind, the information gathered may be more or less useful depending on how fine-grained it is and how far the test content is aligned with the teaching and learning context in question.

Test Development

Construct Definition

Once the use of the test and the types of decisions that are to be based on test scores have been determined, it is then up to test developers to define what the test is supposed to measure. This step should occur before the test itself is constructed. In the context of a language program, it is the responsibility of the test developer (in many cases, the teacher) to make sure that the test is adequate and appropriate in gathering information about the learners (e.g., their level of proficiency or competency). The targeted ability in question is the construct that the test is designed to measure. Construct definition is the first step in making sure that test construction is as systematic as possible.

There are a number of different ways in which constructs can be defined. One way is a construct definition based on a theory about language or language learning. This approach is typically taken when test developers are interested in designing a proficiency test, though it may be used for other types of tests as well. Test developers typically define language proficiency in terms of skills (listening, speaking, reading, writing), elements (grammar, vocabulary, phonology), or both, and may or may not integrate certain skills based on the perceived target language use (TLU) domain (i.e., the way or ways in which the language will be used in the context outside the test). Test developers may also define the construct in terms of a syllabus, textbook, or course objectives, and use these sources as a basis for choosing test format and content. Somewhat differently, constructs in standards-based assessment are defined in terms of teaching or learning standards, or both, which are then used to target abilities on a test. Although syllabi, textbooks, objectives, and standards (which are typically used as a basis for construct definition for classroom tests, placement tests, and standards-based assessments) are often

not explicitly linked to a theory of language and language learning, they are ideally informed by one, though this is not always the case. Chapelle (1998) offers a more comprehensive guide for construct definition including a number of different approaches and theoretical considerations for each.

Constructs under measure are often most transparent for test users in tests with performance-based tasks, where a rubric is used. For instance, if test takers are given ratings on a speaking test based on their performance with respect to grammatical accuracy, meaningfulness, organizational competence, and sociolinguistic competence, speaking ability is being explicitly defined in terms of these components. Therefore, when test developers are interested in including certain domains on a scoring rubric, they need to be mindful that these criteria represent the construct measured on the test. If there is a mismatch between what the test developer perceives as the construct being measured and what is actually being given a score on the test, the validity of the inferences and decisions being based on the test scores may be called into question. Compare this with multiple choice (MC) and limited production tasks where test users are often provided with no explicit representation of what is being measured on the test. In this case, it can be more difficult to see any explicit connection between the tasks on the test and the construct underlying it.

In order to trace how construct definition informs test development, take the example of academic writing for graduate students. In order to create a construct definition, a test developer would first need to ask the question: What is academic writing for graduate students? In other words, what does academic writing look like at the graduate level? What are the characteristics of this type of writing? Where is there potential for variation among writers? Since second language writing is the point of focus here, the test takers' control over grammatical accuracy and complexity, the sophistication and range of vocabulary used, the formality of the tone, and word choice, among other considerations, may be important to the measure of academic writing ability. The organization of the writing will most likely be of concern, as well as the development of the topic and the coherence of the ideas expressed. So, the construct definition might be that academic writing ability can be explained in terms of language knowledge, organizational knowledge, and topical knowledge. This is by no means an exhaustive list, but these are typical considerations used in a definition of academic writing. It is important to remember that a construct definition is simply defining what the target of measurement is—it is not the measurement itself. A test developer would still need to figure out how academic writing ability would best be measured for their purposes. Specifically, the test developer would first need to outline the ways in which academic writing is used in the university context before creating the actual test. This is where the TLU domain comes in. (See Chapter 46, Defining Constructs and Assessment Design.)

Defining the TLU Domain

Before test construction can begin, test developers need to first answer questions about where the test takers are ultimately going to be using the language—be it in an English-medium academic context, in an ESL or English as a foreign

language (EFL) environment, in the workplace, solely in a language-instructional context, or some combination of these. Outlining the types of language use tasks in the TLU domain ultimately helps test developers determine the types of tasks the test takers should be asked to perform on a test. In other words, test tasks are ideally drawn from or based on real-life language use tasks that the test takers need to perform in the TLU domain. For example, if learners in an academic English course will ultimately be using the language in an English-medium university environment, the TLU domain will be primarily academic, including the language used in the classroom, office hours, and formal meetings, but it may also include the language used during more informal interactions, such as social events. Thus, on a test, perhaps these learners will be asked to perform a variety of tasks, including summarizing a lecture, giving an opinion about an article, asking a professor for an extension, or inviting a friend to a discussion group, that are a reflection of the TLU domain. If the course were specifically focused on academic writing as in our example above, perhaps the TLU domain would more narrowly include specific types of writing seen in an academic context, such as essays, research articles, conference papers, annotated bibliographies, technical reports, and critical reflection papers. Even though these different types of writing all tap into the aspects of our academic writing construct (i.e., language knowledge, organizational knowledge, and topical knowledge), there may be variability in writer performance with respect to these elements depending on the type of writing they are asked to perform. Therefore, it is crucial that test developers have a clear idea of which language use tasks in the TLU domain elicit the most representative sample of the test takers' ability, so that, when it comes time to construct the test, the developers are able to choose test tasks that provide the best information about what the test takers know and can do. In this case, out of the many TLU tasks that learners may perform in a real-life university context (e.g., essays, research articles, conference papers, annotated bibliographies, technical reports, and critical reflection papers), given time and resource constraints, a test developer will probably need to select one (or two) types of academic writing tasks to include on a test. More than likely, a test developer will choose an academic essay for the test since the other types of academic writing require too much research or topical knowledge for a testing context. However, it is still important to bear in mind the importance of implementing a systematic framework of test specifications when constructing the actual test, even if a test developer has a basic idea of what the test will look like.

Test Specifications

Once the TLU domain is defined, test developers can begin the process of test construction. When designing an assessment, details about the test format and content need to first be outlined in a systematic way. Using a systematic process to create a framework within which to develop test tasks ensures that consistency of measurement (i.e., reliability) is maximized and unwanted variability due to the test method is minimized. This procedure entails designing specifications for the test and test tasks. Creating test specifications within a framework also allows for parallel forms of a test to be more easily created if that is the need of the

stakeholders. In the context of a classroom, specifications can help to ensure that the test correctly relates to a teaching syllabus or other features of the teaching and learning context. In a more high stakes situation, specifications are important because they bolster test quality and help to demonstrate that the decisions based on test scores are fair and valid. Finally, test specifications, and task specifications more specifically, can be used to link the test tasks to the TLU domain, which will help ensure a precise measure of the learner's language ability in a given context.

Bachman and Palmer (1996) provide a comprehensive framework of specifications that includes both the test as a whole and the characteristics of the test tasks contained within the test. With respect to test characteristics, test developers need to specify the characteristics of the test setting (e.g., participants, location, and time of the test) and the characteristics of the test itself (e.g., overall test instructions, test structure, time allotment, any cut scores for the test, or weighting of test sections). Test developers need to also outline a number of characteristics of the individual tasks within the test as well. These include specifying the individual task instructions, the format, language, topical and strategic characteristics of the input and expected response, scoring method, and the relationship between the input and expected response. Once these elements have been specified, the actual test tasks can be written. (See Chapter 47, Effect-Driven Test Specifications.)

Task Types

There are two main classes of task type that test developers need to know: selected response and constructed response. Selected response tasks include conventional MC (gap-fill, sentence completion, etc.), matching (fill-in with lists), and discrimination (true/false, same/different, etc.). The second class, constructed response, can be further subdivided into limited production and extended production task types. Limited production tasks typically involve brief, written responses, including short answer, fill-in-the-blank, cloze (examinee is asked to fill in several blanks within a passage), and discourse completion tasks (DCTs) (examinee is asked to provide lines of text to complete lines in a dialogue). Extended production tasks typically involve longer responses (either written or spoken), such as structured question or information gap tasks, stories, reports, essays, interviews, role plays, and simulations. The selection of a particular task type will depend on the nature of the information that stakeholders need to get from tests. There are many sources of practical information on different task types (e.g., Hughes, 2003; Coombe, Folse, & Hubley, 2007) and some are even tailored to certain skills (e.g., speaking) or specific populations (e.g., K-12 learners). (See Chapter 52, Response Formats.)

In an ideal world, there exists an alignment between the instructional tasks (if the test is to be given in the context of instruction) and the test tasks, and also an alignment between the test tasks and the real-life tasks the test takers will be asked to perform outside of the test. Now, this *authenticity of task* is achieved through the test tasks being a close approximation of the TLU tasks (Bachman & Palmer, 1996); however, sometimes stakeholders are simply interested in gathering information about, for instance, the test takers' knowledge of the past perfect or article usage and, therefore, may develop a quick, MC test. Though perhaps inauthentic,

this type of test may provide precise and sufficient information for the stakeholders with minimal negative impact in terms of time, resources, and test-taker affect. In other words, no task type is inherently bad. It just depends on the type of information that the stakeholders are interested in gathering. Ultimately, test developers should try to maximize authenticity of task (i.e., create tasks that reflect real-life situations) in their quest to capture the TLU domain in their test tasks, while still being mindful of the effect that task type may have on practicality considerations on the one hand and the types of inferences that can be based (or not) on test performance on the other.

Item and Task Writing

The type of item or task that is selected should be a function of the desired outcome (e.g., a learner's ability to comprehend a listening passage) that is being tested (Lane, Haladyna, Raymond, & Downing, 2006). Whether tapping into receptive skills (e.g., reading and listening) or productive skills (e.g., speaking and writing), test developers should design task types that result in the most adequate means of capturing aspects of a learner's language ability. Some task types require expert judgment, necessitating human raters (or machine scoring or both) to evaluate the learners' performance, while other tasks are scored objectively and require no expert judgment (e.g., MC items). Choosing the appropriate item or task format depends on what type of information is needed about the performance of test takers and what decisions are to be made about them. Again, although it is preferable to maximize authenticity of task to have the test reflect the domain in which the test taker will use the language, less authentic item types (e.g., selected response) are often preferable when teachers want to assess, for example, learners' knowledge of grammar (e.g., past conditional), their comprehension of a reading passage (e.g., "What is the best title for the article?"), or their comprehension of a radio program (e.g., "What does the man say is the most current threat to the economy?"). Although selected response items are easily scored, writing items that perform well requires extensive and extended training (Bachman & Palmer, 2010).

Whether writing selected response or constructed response items, there are a few general points to follow. Test items should try to (a) include instructions that are clear, concise, and elicit appropriate responses; (b) tap into a testing point that is connected to the test construct or an instructional objective; (c) follow standard conventions for grammar, punctuation, and spelling (of a particular language variety); (d) include clear and unambiguous language within the item, which helps avoid its being tricky or unanswerable with the background knowledge of the examinees; and when possible, (e) include an example item to minimize ambiguities, especially in the event of introducing a new item type.

Instructions should be given for each task. For selected response tasks in particular, they should be short, concise, and unambiguous, but still provide enough information so that the test taker will be able to fulfill the task. Instructions should elicit the desired output and minimize anything unrelated to the construct. Each item is composed of a "stem," which is usually a one-line question or statement (e.g., "What is the best title for this passage?") or a sentence or dialogue with a

Table 45.1 Anatomy of an MC item

Choose the best answer to complete the sentence.		Instruction line(s)	
Jack:	How _____ you?	Stem (in the form of a dialogue)	
Sara:	I'm fine. Thanks.		
	a) is	Distracter	4 options
	b) be	Distracter	
	c) are	Key	
	d) being	Distracter	

Table 45.2 General rules for MC item writing

What to do	*What to avoid*
Measure a single testing point (e.g., write an item measuring tense only rather than tense and word meaning together).	No option should cue another; keep items independent of one another.
Create distracters that are plausible and attractive. Avoid illogical distracters.	No item should have more than one key (correct answer).
Use vocabulary and grammar consistent with the test takers' level of understanding.	No option should "stick out" from the others. Item options should look like a coherent set.
Employ a similar level of grammatical complexity when writing options.	Avoid negative forms in the stems and in the options, when possible.
Write options that are similar in length.	No option should cancel another one out. Avoid using words like "always" and "never."
Include all necessary information in the stem (e.g., if words are repeated in the options, move them up to the stem).	Avoid creating an item that taps into more than one testing point (e.g., word meaning and morphosyntactic form together).

fill-in-the-blank. Stems need to be accurate and contain only the necessary information to target the testing point. Extraneous information will only require more reading on the examinees' part, which might detract from what is the object of measurement. Most MC questions have three or four options, composed of one key and three distracters. The "key" is the correct answer, which should unequivocally be the best answer among the options. Ideally, the three (or two) distracters are equally "distracting," but this rarely occurs. Typically, only one or two distracters are chosen by examinees at lower levels of ability. Table 45.1 exemplifies the anatomy of an MC item.

Before attempting to write test items, it is important to be aware of what to do and what to avoid. Table 45.2 outlines some general rules to keep in mind.

Since writing successful MC items is often challenging, test developers should reconsider using MC items unless there is a process of item analysis, whether qualitative or quantitative, that aims to evaluate the test content. Seeking the help of qualified experts to look at the content of the test and the items themselves can be helpful, but the item review process should be as systematic as possible, including, at a minimum, the use of item-writing checklists.

As an alternative, constructed response tasks, when designed well, can provide stakeholders with a great deal of information about test takers. Since more learner output is encouraged in this item type, there are more opportunities for the stakeholders to directly view the evidence of what the test takers can do in a given task. The challenge comes in honing the prompt to elicit the targeted response from the test taker. For example, prompts that are too generally worded will result in wide variability in test takers' responses. More narrowly focused prompts will lessen ambiguity and will help focus the test takers into providing the desired language, structure, or both for a given task (Bachman & Palmer, 2010). (See Chapter 48, Writing Items and Tasks.)

What is most important in writing items and tasks is that they adhere to the test specifications as written by the test developer, since they refer back to the construct or instructional objectives. Many testing organizations compose elaborate item-writing guidelines or checklists, which can provide indispensable guidance for novice and experienced test writers alike. Documents such as these typically include example items that help add clarity and purpose to a test writing session. Test developers working with a team of test writers should consider creating a set of guidelines for a given test. Finally, asking a colleague who does not have familiarity with the test takers to give feedback on a newly revised test can also be a valuable exercise (Davidson & Lynch, 2001).

Ultimately, no matter what kind of items or tasks are used, issues of test fairness must always be addressed so that stakeholders can determine whether the difference in examinees' test performance involves factors that are related or unrelated to the examinees' true language ability. Kunnan (2004) creates a Test Fairness Framework in which he suggests how test developers can make mindful decisions about possible systematic bias related to (a) dialect, content, and topic, and (b) group performance (e.g., gender, age, language group, etc.). When such biases are identified, Kunnan recommends flagging these items for a thorough content review, from which decisions about reviewing, modifying, or deleting items can be made. (See Chapter 66, Fairness and Justice in Language Assessment.)

Test Administration

Although test developers may not necessarily be the administrators of a test, they, too, need to consider a number of variables that may affect test performance, and how these things relate back to test development. If the test administrator is also the test developer, they likely have even more information at their disposal to minimize unwanted variability in scores due to the administration process. First, the test environment should be comfortable and free of unwanted distractions (e.g., construction noise outside a classroom during a listening test). Second, test takers will feel more prepared if the format and content of the test tasks are familiar (e.g., ones similar to those they have previously encountered during classroom instruction) or the tasks reflect the TLU domain. Third, test takers should be given access to as much information about the test as possible without unfairly advantaging some test takers over others (or giving them the answers, obviously). Practices such as mock exams and giving out copies of the rubric well in advance

can maximize test-taker performance. Transparency is key in helping the test takers feel prepared, relaxed, and focused during a test.

Some test developers are interested in obtaining feedback, either before, during, or after the test administration, in order to improve a test for future administrations. This feedback may include information about the test administration or the test items or tasks themselves. It is important to keep in mind that asking the test takers to complete a checklist, questionnaire, or interview during a live administration of a test can induce anxiety in some. Although it is possible to minimize the negative impact of such information-gathering techniques, it is preferable to obtain feedback during a pilot version of the test. (See Chapter 53, Field Testing of Test Items and Tasks.)

It is also important to note here that despite efforts made by test developers to adhere to a particular set of guidelines and specifications, an influx of test preparation courses have cropped up in recent years on a global level. These test preparation courses feature a range of test-taking strategies intended to increase examinees' level of test-wiseness, including making efficient use of time and guessing. While some test preparation courses equip examinees to prepare for the content of the test (e.g., speaking ability tasks), some less ethical courses prepare them to become familiar with the idiosyncratic characteristics of a test developer, making the score results of the examinees questionable. In other words, have the examinees reached a level of proficiency (or achievement, mastery, etc.) or were they merely using their ability to decode or "game" a test (test-wiseness) according to their knowledge of item construction patterns of a particular test? Such issues make it difficult to make genuine and informed decisions about an examinee's true language ability. Ideally, this is not something most stakeholders will encounter, but it is certainly something to be aware of. (See Chapter 68, Consequences, Impact, and Washback.)

Scoring

Scoring Methods

There are a number of different scoring methods that all test developers should be aware of. The process of test scoring obviously comes after the test has been administered (or during the test, in some cases), but test developers should be thinking about what kind of scoring procedures would be most useful and beneficial when designing and operationalizing the test constructs into test tasks. Specifically, the test developer (and possibly other stakeholders as well) should identify the type of information that is needed and also how detailed that information needs to be in order for the best possible inferences and decisions to be made in a given context. In some cases, coarse-grained information, such as a total score, will be sufficient; in other situations, stakeholders will want highly detailed, fine-grained information on which to base their decisions. Thus, selecting the appropriate scoring method is crucial, particularly in high stakes situations.

There are three main types of scoring methods: right/wrong scoring (with one or more criteria for correctness); rating scales with limited production items

(which are typically limited in both the number of scale levels and also the level of detail in the descriptors); and holistic or analytic rubrics with extended production items (which are typically broader in both scale length and the level of detail in the descriptors). Right/wrong scoring is used when test items can be scored as either "right" or "wrong," typically on one dimension. Conventional selected response (e.g., MC) items are typically scored right/wrong, or dichotomously. In this case, when a response is considered right, or correct, it is usually given a score of "1"; responses that are wrong, or incorrect, are given a score of "0." Dichotomous scoring is the most straightforward to implement, but provides the least detailed feedback for the test users when compared to other methods. Test responses can also be scored right/wrong on more than one dimension. For instance, a limited production item (e.g., cloze) may require the test taker to produce both the correct form (e.g., *had run* as opposed to *ran*) and the correct meaning (e.g., *run* as opposed to *walk*) of a particular verb in a blank. The response can be scored as either correct or incorrect on two dimensions (i.e., correctness of form and correctness of meaning). In this case, test takers could be given *two* scores for each blank in the cloze test—each score being dichotomous in and of itself. In this case, right/wrong scoring with multiple criteria for correctness would be being implemented.

The second type of scoring involves rating scales. Rating scales are typically used in scoring limited production items. In this case, the test taker is producing more than single words or phrases; therefore, there may be degrees of correctness on one or more dimensions, rather than being right or wrong. For example, test takers are asked to complete a conversation as part of a DCT by producing one or two short sentences. Responses may be fully precise and meaningful or full of errors and incoherent, but they may also be somewhere in between. In this case, the test developer may decide that responses should be scored on a continuum rather than simply right/wrong. Therefore, individual responses are scored for grammatical accuracy and meaningfulness, each on a scale of 0–3. It is up to the test developer to decide what is being operationalized, elicited, and feasible for scoring in a given task (e.g., perhaps pragmatic appropriateness is also elicited). Using rating scales provides more information for test users than does right/wrong scoring, but it is typically more coarse-grained than information about test takers obtained through the use of scoring rubrics, since there are usually more score bands and detailed descriptions of behavior associated with rubrics.

The third type of scoring, using rubrics, is typically associated with extended production items or tasks, such as in performance-based speaking or writing assessments. Since extended production responses contain a great deal of language, right/wrong scoring or relatively simple rating scales are often insufficient to capture the heterogeneity along several, potentially distinct, dimensions of test-taker performance (e.g., organizational competence, topic development, and language control in a compare-and-contrast essay). In addition to accounting for several domains of knowledge or ability, rubrics allow for these domains to be scored on multiple bands, or levels. Each level ideally has a detailed description of what the performance looks like at that particular score band, and the descriptions should use parallel language in all bands (e.g., adverbs of frequency, comparative adjectives, etc.). These descriptors help raters identify the characteristics

of a particular performance and link it to the appropriate score. This alignment helps objectify the rating process and maximize accountability. Holistic rubrics collapse all domains under measure (and their associated descriptors) within one large scale. With holistic scoring, test takers receive a single score for their performance. By contrast, analytic rubrics separate each knowledge or ability component into its own separate scale, thus giving test takers as many scores in their analytic score profile as there are domains in the rubric. Quite obviously, analytic scoring provides more fine-grained information about test-taker performance, but it is usually more labor intensive in terms of rater training and the time required for scoring. Ultimately, it is up to the test developer to decide which type of scoring, holistic or analytic, is most appropriate given the type of information the stakeholders require. However, no matter which type of scoring is chosen, it is important to remember that the construct under measure should always be reflected in the domains and descriptors in the scales. (Sess Chapter 51, Writing Scoring Criteria and Score Reports.)

Raters and Rater Training

Training raters to score written or oral test samples is an important step in the test development process. Providing solid rater training contributes to reliable and valid interpretations of examinees' scores on test tasks that elicit extended learner output (Cizek & Bunch, 2007). For constructed response tasks (e.g., a persuasive essay or an oral task), raters are needed to score the examinees' written or spoken performance. Ideally, ratings should be blind to reduce bias and raise the notion of fairness. Ultimately, it is important for raters to have a strong knowledge base in matters related to scoring procedures. This is achieved through (a) systematized training, (b) well-defined, unambiguous benchmarks, and (c) a precise rubric whose descriptors depict the domains (i.e., match the construct) of a given task. Without these, raters will fail to agree on what constitutes superior performance on a given task.

The first step in rater training is to have the raters methodically follow standard setting procedures (Cizek & Bunch, 2007). The goal of a standard-setting activity (often called a norming session) is to train raters to be consistent (and equally severe) in their ratings. Before a standard-setting session takes place, it is common for raters to be given training materials that provide examples of different levels of performance that are identified as being representative of each band of the rating scale. During the actual standard-setting session, an impartial arbiter (e.g., the test developer) provides the panel of raters with writing samples that they are asked to score. Raters then compare results in a substantive discussion about the examinees' individual performance. As long as the descriptors are adequate, raters should not be more than one band apart. However, when disagreements arise, an adjudication process occurs in which discrepant raters are asked to provide rationales for their scores using the rating scale descriptors to support their argument, and differences are usually negotiated to reach a "normed" score across the raters. Even if raters appear normed after a standard-setting activity, it is important that they are re-normed periodically to maximize consistency. (See Chapter 80, Raters and Ratings.)

Reporting Test Results

The first thing to consider when thinking about reporting test results is the audience. Are they test takers, teachers, administrators, parents, employers, governments, or some combination of stakeholders? Different types of stakeholders often require different types of information, and making the information transparent and accessible is key. Therefore, it is up to the test developer to create test tasks to elicit the targeted information, and also choose scoring procedures (e.g., holistic versus analytic scoring) that let the grain size of the information needed be made available for reporting. What would be most helpful given the context? Perhaps for some stakeholders, a single numerical score is sufficient for their purposes (e.g., meeting a cut score for university admissions). However, for other stakeholders, detailed, diagnostic information may be indispensable for prescribing future teaching and/or, learning, or both, as in the case of an English for academic purposes program. Of course, diagnostic feedback is not always practical to give (or, for that matter, available), but since many test takers are also learners, providing as much feedback as possible can prove beneficial to them and their future learning, and hopefully enhance the positive impact of the test. (See Chapter 58, Administration, Scoring, and Reporting Scores.)

Test Data Analysis

Test developers may or may not be the individuals responsible for analyzing test data, but understanding the most commonly used approaches is important for completing the chain of test development. Since data analysis provides insight into the psychometric properties of a test, research findings often lead to subsequent iterations of test revision, and hopefully, improved versions of a test. As part of this process, test developers may again be called on to make changes to test specifications, items or tasks, test administration, scoring procedures, or a combination of these. At a minimum, test developers should understand the purpose of various statistical analyses, what kind of information they provide, and how this information relates back to future test development decisions. If test developers are interested in a more in-depth study of these analyses, Bachman and Palmer (2010) provide a treatment that is specific to language testing, and Carr (2011) provides a tutorial for analyzing language test data using Excel. (See Chapter 56, Statistics and Software for Test Revisions; Chapter 69, Classical Test Theory.)

Descriptive Statistics

Descriptive statistics provide measures of central tendency and dispersion. Central tendency refers to how well (or how poorly) the broad middle of the test takers performed. Knowing about where our students are "on average" can help inform teaching, learning, and future test revisions (i.e., was the test easier or harder than expected? If so, how can teaching, learning, or testing be changed?). Typically, central tendency is described using the mean, median, or mode. The

most commonly used measure of central tendency, the mean, is the average score on the test. Similar to the mean, though not interchangeable, is the median. The median is the middle score on the test. In other words, if you physically ordered the test papers from lowest score to highest score and then found the test in the middle of the pile, that score would be the median. Since the procedure for obtaining the mean involves an averaging process where even very high (or low) test scores become part of the numerical calculation, the median is not as sensitive to outliers, because the ordering of the test scores in terms of highest, second highest, third highest, and so forth, does not take into account the magnitude of the differences between the scores. Therefore, in cases where outliers' scores skew the mean to be either significantly higher or significantly lower than the true representation of the broad middle test takers' performance, the median may be a better indicator of central tendency than the mean. The mode, or most commonly occurring score, also provides information about the central tendency of the group. Bimodal, or even trimodal, distributions may be seen when there are distinct subgroups of test takers within a test-taker population (e.g., heritage language learners, ESL vs. EFL learners, etc.). If any of these measures of central tendency indicate a different result from what was expected of the average, middle, or most common performance, this may be an indication that test developers need to revisit the design of the test.

Dispersion tells us about the variability in the test-taker population. Are the test takers similar to one another, or very different? In other words, do the results indicate a homogeneous population, and thus one whose members' needs can be addressed in a similar way? Or is it heterogeneous, and, therefore, may we need to contend with a host of diverse teaching, learning, or testing issues? There are two principal measures of dispersion: range and standard deviation. The range is the interval between the lowest and highest scores on a test. Though potentially useful if the population is very large, the range is sensitive to sample size, and thus may be misleading if there are outliers in the data. For example, if most test takers receive scores in the 90s or 100 out of 100, but there is one test taker who receives a score of 10, the range will encompass nearly the entire score spectrum (i.e., 90 points) even though only one test taker received such a low score. In this case, the range of scores on the test would not really be a good reflection of the variability (or the central tendency) of the group. In contrast, the standard deviation is typically a better indicator of how much variability there is in the test scores, since it is an average of how much all the scores deviate from the mean. A high standard deviation would be an indication of a lot of variability in the scores (i.e., heterogeneous population, platykurtic distribution), whereas a low standard deviation would be an indication of very little variability in the scores (i.e., homogeneous population, leptokurtic distribution). A score distribution can also be skewed in such a way as to indicate that test takers did either better or worse than the mean. In other words, the test was relatively easy or hard in terms of probability. Score distributions that are negatively skewed show that a test was relatively easy, and a positive skewness indicates relative difficulty. Classroom achievement test scores are often expected to show negative skewness, whereas a pre-unit check or diagnostic assessment might show positive skewness. Again, if the results are unexpected, this may be an indication that the test developer needs to revise the test

(e.g., perhaps add more difficult or more easy items) so that the distribution of the scores matches the expectations of the measurement for future administrations.

Reliability Analysis

In order for test developers to determine the utility of an assessment, the reliability, or consistency of measurement, must be determined to show that the test results are trustworthy. If very different test results are obtained when any number of different conditions of the assessment vary (e.g., items or tasks, occasions, raters, forms, or a combination of these), the quality of the information obtained can be considered untrustworthy. Ideally, test scores (i.e., observed score variance in measurement terms) are a close approximation of the test takers' knowledge or ability (i.e., true score variance in measurement terms). The better the test scores represent the test takers' true ability, the higher the reliability will be. However, since measurement is never without error and reliability is never perfect, test results are expected to vary somewhat, but it is up to the test developer and other stakeholders to determine the level of consistency that is acceptable in a given context. Obviously, the higher the stakes of the test, the more important it will likely be for a high level of reliability to be obtained.

Reliability can be maximized through systematic test development procedures, but it is not until the data analysis phase that reliability can be statistically determined. A number of different reliability estimates can be obtained, one internal and three external. Arguably the most important type of reliability estimate is internal consistency reliability. Internal consistency reliability, usually measured with Cronbach's alpha, gives an estimate for the extent to which a test score (i.e., observed score variance) reflects the test takers' theoretical score (i.e., true score variance), rather than error. The better this correspondence is, the higher the reliability will be. There are three different types of external reliability that can be estimated. First, test–retest reliability can be calculated when the same test form is administered on multiple occasions. This type of reliability indicates the extent to which the test taker gets the same score from administration 1 to administration 2. Second, parallel forms reliability can be obtained when one or more forms of a test are given. This type of reliability indicates the extent to which scores on one form of a test are comparable to scores from another test form (created with the same set of test specifications). Last, rater reliability, though external to the test itself, is often calculated as another indication of the consistency of measurement. Inter-rater reliability can be calculated when two or more human raters assign scores to test performance samples, as in a writing or speaking assessment. A high inter-rater reliability would be an indication that the raters are assigning similar scores. Intra-rater reliability can be calculated as an alternative to inter-rater reliability, when a single rater (as opposed to more than one rater) is assigning two or more scores to each test performance. Multiple ratings of a test performance by a single rater sometimes occur, for instance, when a teacher is solely responsible for rating students' work, and would like to rate each student's test twice to minimize any ordering effect. Which reliability estimates are calculated will ultimately depend on the format of the test, the context of the test administration, and scoring procedures used. (See Chapter 70, Classical Theory Reliability.)

Item Analysis

Once data have been collected from a test administration, item-level information can be used to modify, and hopefully improve, a test. Statistics that indicate item difficulty (also known as item facility) and item discrimination are commonly calculated by test researchers as a first step in item analysis. Item difficulty for dichotomously scored items is usually given in the form of a *p*-value, which is the proportion of test takers who answered an item correctly divided by the number of test takers who answered the item (which is usually the same as the number of test takers who took the test). *P*-values (which are equivalent to item means for dichotomously scored items) range from 0 to 1, with values closer to 1 indicating very easy items and values close to 0 indicating very difficult items. Item difficulty for items not scored dichotomously would be determined simply by calculating the mean for an individual item or task, and would not necessarily range from 0 to 1. The values that would indicate difficulty or ease of an item or task would be interpreted relative to the scale on which the items or tasks were scored and the expected performance of the test takers given the context. Item difficulty can be examined to determine the extent to which the difficulty of the items meets the expectations of the measurement. If the items are too difficult (or too easy), stakeholders, who can include the test developer, may decide that certain items, or the test as a whole, need(s) to be revised.

One goal in testing is often to separate test takers from one another in terms of their knowledge or ability (e.g., masters from nonmasters). Thus, test developers ideally create items that high ability test takers are able to answer correctly more often than low ability test takers. Therefore, another item-level statistic, item discrimination, is useful for determining how well an item performs in terms of separating high ability test takers from low ability test takers. Item discrimination (for dichotomously scored items) is typically calculated using a point biserial correlation. Values above .3 indicate that an item is effectively separating the test takers in terms of their ability, and can be retained as is. Values of .2 to .3 indicate borderline effectiveness and potential for item revision, and values lower than .2 show strong evidence that an item needs to be revised or deleted from a test. Item discrimination values below 0 indicate that lower ability test takers performed better on a given item than did higher ability test takers. Since this finding runs completely counter to expectations, any item showing negative discrimination should first be examined for miskeying, double-keying, potentially confusing language in the instructions or in the item itself, or a combination of these, before it is revised or ultimately rejected. It is important for test developers to keep in mind that any time an item is removed from a test, item discrimination statistics for the other items will change slightly. Therefore, it is best to perform these calculations again for each iteration of the analysis and subsequent test pilots, since item-level statistics tend to become more stable as improvements to a test are made. Finally, once items (and their associated statistics) can be considered acceptably stable for a given context, an item bank can be constructed so that test developers can pick and choose from a pool of items of varied difficulties to create new forms of a test. As part of the item bank, statistics such as item difficulty and discrimination can be catalogued and revised each time an item is administered.

Not only is item banking useful for testing companies in terms of cost savings, but also classroom teachers may find item banking useful either for themselves when teaching the same level year after year, or to share with colleagues who may be teaching in the same program. (See Chapter 49, Item Banking.)

Higher-Level Analyses

Even if test developers are not the ones responsible for performing the actual analysis of test data, it is important that they have an awareness of the higher-level analyses that are available to researchers, since test revisions, (re)piloting, and changes to scoring procedures that often occur as a result of research findings usually become the responsibility of the test development team (and administration personnel). The most commonly used statistical models in language assessment research are item response theory (IRT), structural equation modeling (SEM), generalizability theory, and their related models. The specific statistical model that is employed will depend on the types of questions the stakeholders want answered. A comprehensive treatment of these models is beyond the scope of this chapter; however, information relating to each that is also specific to language assessment can be found in Chapter 72, The Use of Generalizability Theory in Language Assessment; Chapter 73, Exploratory Factor Analysis and Structural Equation Modeling; Chapter 75, Item Response Theory in Language Testing; and Chapter 77, Multifaceted Rasch Analysis for Test Evaluation.

Conclusion

The purpose of this chapter is to provide the fundamental concepts that anyone developing a language test needs to consider. Oftentimes in a context where there are limited resources, one person (e.g., a classroom teacher) is tasked with having to develop a test for the assessment of their learners. Having an understanding of the basic concepts of test construction is obviously critical, but also having working knowledge of test administration practices, scoring procedures, and data analysis is important since any and all of these elements of the assessment process can directly affect test development. This chapter provides an overview of the conventional uses of tests, followed by instruction in construct definition and, ultimately, training in test construction, test administration, scoring considerations, and data analysis. The systematicity with which such processes are conducted can greatly affect the quality of a language test.

References

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests.* Oxford, England: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford, England: Oxford University Press.

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

Chapelle, C. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman & A. D. Cohen (Eds.), *Second language acquisition and language testing interfaces* (pp. 32–70). Cambridge, England: Cambridge University Press.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage.

Coombe, C., Folse, K., & Hubley, N. (2007). *A practical guide to assessing English language learners*. Ann Arbor, MI: Michigan University Press.

Davidson, F., & Lynch, B. K. (2001). *Testcraft: A teacher`s guide to writing and using language test specifications*. New Haven, CT: Yale University Press.

Hughes, A. (2003). *Testing for language teachers* (2nd ed.). Cambridge, England: Cambridge University Press.

Kunnan, A. J. (2004). Test fairness. In M. Milanovic & C. Weir (Eds.), *European Year of Languages conference papers, Barcelona, Spain* (pp. 27–48). Cambridge, England: Cambridge University Press.

Lane, S., Haladyna, T., Raymond, M., & Downing, S. M. (2006). *Handbook of test development*. Mahwah, NJ: Erlbaum.

Suggested Readings

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge, England: Cambridge University Press.

Haladyna, T. M., & Downing, S. M. (1989). A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education*, *2*, 37–50.