CHAPTER 18

CONSIDERATIONS FOR ACHIEVEMENT TESTING OF STUDENTS WITH INDIVIDUAL NEEDS

Rebecca Kopriva and Craig A. Albers

UNCORRECTED PROOFS © AMERICAN PSYCHOLOGICAL ASSOCIATION

This chapter summarizes the primary issues and topics germane to constructing defensible large-scale academic assessment systems that are accessible for two groups of students, English learners (ELs) and students with disabilities (SwDs). Most of the research and policy to date has occurred within the context of large-scale statewide summative achievement tests designed to be administered yearly to students in Grades 3 to high school, and this material is the bulk of what is summarized in this chapter. Much of what is discussed can be generalized to college-level tests and psychological assessments as well, and to other types of K–12 achievement assessments such as benchmark tests (tests that occur over schools several times during the academic year and are meant to gauge the partial-year performance of students) and standardized formative assessments (these can be course-embedded assessments or other stand-alone assessments designed to support instruction). Although the purposes of the summative, benchmark, and formative achievement assessments differ (they range from evaluation of status knowledge and skills to providing more fine-grain information teachers can use to adapt instruction or clarify misunderstandings), issues of access are usually similar. As access might be adjusted for the different types of achievement testing, this will be noted when possible. Furthermore, of course, the notions about how to properly evaluate the learning of young K–2 children with special needs is examined to a limited extent. To date, this is an important area that, unfortunately, has received limited attention.

Although the chapter does not address the assessment of English language proficiency for ELs, emergent skills in the development of English certainly affect how ELs need to be asked questions about academic content and the proficiency and strategies they need to have in demonstrating their understandings using appropriate communication supports (see Chapters 10 and 17, this volume). For SwDs, this chapter does not examine cognitive or emotional assessment considerations (see Chapters 3 and 6, this volume). Also not specifically addressed is the issue of ELs who have also been identified as having a disability (see Chapter 9, this volume).

What does it mean to make academic assessments and especially academic assessment systems accessible? This is a complex question, addressed in some depth by Rigney, Wiley, and Kopriva (2008) and then examined in detail throughout Kopriva (2008). Recently, Winter (2010) has used the lens of test score comparability to highlight various considerations that need to be resolved when access in achievement systems means deciding under what conditions and with what evidence can scores be considered interchangeable when students are taking the same or similar tests under different conditions.

In essence, the goal of achievement tests is to be able to appropriately capture the knowledge and skills intended by the test, and more specifically, by each item, question, or task. Accessibility within this

The contents of this chapter were developed in part by Enhanced Assessment Grants (S368A090029) and (S368A080004) from the U.S. Department of Education and awarded to Rebecca J. Kopriva and Craig A. Albers, respectively.

DOI: 10.1037/14049-018
APA Handbook of Testing and Assessment in Psychology: Vol. 3. Testing and Assessment in School Psychology and Education,
K. F. Geisinger (Editor-in-Chief)
Copyright © 2013 by the American Psychological Association. All rights reserved.

APA-HTA_V3-12-0603-018.indd 369 12/10/12 1:30 PM

context means that students will be able to properly hear and respond to the intent or target of each question. Any time a question is presented to the student, methods of communication are employed—these methods are meant to facilitate the interaction of the targeted question and response between test maker and test taker. These methods are ancillary to the intended meaning of the task, and sometimes, instead of facilitating the interaction, they act as barriers that wholly or partially prevent the transference of the question or the intended knowledge and skills between the test taker and the test maker. There seem to be three aspects of access that occur within each achievement task. First, the student must have sufficient access to how the meaning and the requirements are conveyed in the task—in other words, what is the task or question asking? Second, for the student to initiate and sustain problem-solving activities relative to the task requirements, students must be able to access their procedural skills and other content assumed by the task and must have the tools necessary to implement the activities. Third, students must be able to access their representation skills commensurate with the representation constraints in the particular task. This means the task or question must be set up in such a way that the student can adequately convey their skills or knowledge to the test maker.

To be able to address accessibility adequately, an understanding of the two populations is crucial. A brief demographic summary, outline of some assessment related challenges, and a policy overview of each of these groups are described in the following sections.

Students With Disabilities

Demographics and assessment-related challenges. More than 6.5 million infants, toddlers, children, and youth have been identified as exhibiting specific developmental delays or meet criteria for at least one of the designated disability categories under the Individuals With Disabilities Education Improvement Act of 2004 (IDEA, 2004), depending on their age (U.S. Department of Education, n.d.). Part B of IDEA pertains specifically to schoolchildren and youth. Under Part B of IDEA (2004), children and youth between the ages of 3 and 21, along with their families, are afforded special education and related services upon meeting the criteria of at least one of the following disability categories: mental retardation, hearing impairments (including deafness), speech or language impairments, visual impairments (including blindness), emotional disturbance, orthopedic impairments, autism, traumatic brain injury, other health impairments, and specific learning disabilities. The majority of children and youth between the ages of 6 and 21 who are served under Part B of IDEA are classified as having specific learning disabilities (43.6%), followed by speech or language impairments and other heath impairments (19.2% and 10.5%, respectively; Data Accountability Center,

UNCORRECTED PROOFS © AMERICAN PSYCHOLOGICAL ASSOCIATION

TABLE 18.1

Disability Categories of Children and Youth, Ages 6 to 21 Years, Served Under IDEA Part B

Disability category	Percentage of students
Specific learning disabilities	43.6
Speech or language impairments	19.2
Other health impairments	10.5
Mental retardation	8.3
Emotional disturbance	7.3
Autism	4.3
Multiple disabilities	2.2
Developmental delay	1.5
Hearing impairments	1.2
Orthopedic Impairments	1.0
Traumatic brain injury	0.4
Visual impairments	0.4
Deaf-blindness	near 0

APA-HTA_V3-12-0603-018.indd 370 12/10/12 1:30 PM

2007). The remaining disability categories, along with the percentages of children and youth between the ages of 6 and 21 who are served under IDEA Part B and classified as having such disabilities, appear in Table 18.1 (Data Accountability Center, 2007).

In general, SwDs are included in achievement assessments as they are written and administered, by using accommodations along with the general test forms, through modifications of the general test forms or testing conditions, or through using alternate assessments. The goal of any adaptations is to provide more valid and accurate information about the constructs being measured than would be the case when these students take the general assessments under typical conditions. Test accommodations usually fall under the following categories: presentation accommodations, equipment and materials accommodations, response accommodations, scheduling and timing accommodations, setting accommodations, and linguistic accommodations. In the nomenclature of the educational content testing industry, the term *modifications* of the general test denotes that the modifications affect how the constructs are measured through making changes to test modality, complexity, space, time, language, and possibly other aspects (Poteet, 1990). Alternate assessments are intended to facilitate inclusive assessment for students with significant disabilities and must yield information about students' achievement for purposes of statewide accountability. Ideally, alternate assessments should also provide instructional utility. Each of these is discussed more later in the chapter.

Policy overview. Two significant pieces of federal legislation require that SwDs be included in standardized assessment programs: the IDEA, for SwDs only; and the current authorization of the federal Elementary and Secondary Education Act (ESEA; No Child Left Behind [NCLB], 2001) legislation, for all students in public schools. Both of these laws were designed to improve the academic achievement of all students through high expectations and high-quality education programs.

English Learners

Demographics and assessment-related challenges. Estimates suggest that approximately 25% of all U.S. students currently in schools are ELs (Hernandez, Denton, & Macartney, 2008). Of these, children of immigrants now constitute one fifth of all U.S. school-age children, for which a large majority of the households may be described as *linguistically isolated*, which means that no one in the household age 14 or older speaks English exclusively or very well (Capps, Fix, Murray, Ost, Passel, & Herwantoro, 2005). Furthermore, ELs consistently perform below grade level in all content areas. For instance, on the 2005 National Assessment of Educational Progress (NAEP), 46% of EL fourth graders scored "below basic" in mathematics as compared with only 18% of non-ELs; for eighth graders, 71% of ELs scored below basic as compared with 30% of non-ELs (Perie, Grigg, & Dion, 2005); achievement gaps between EL and non-Hispanic White students were 35% in Grade 4 and 50% in Grade 8 (Fry, 2007). ELs are also nearly twice as likely as their native English-speaking peers to drop out of high school (Rumberger, 2006; Silver, Saunders, & Zarate, 2008). Gándara and Rumberger (2009) attributed the higher dropout rate to schools' lack of academic and social supports for ELs beginning well before high school. Callahan and Gándara (2004), among others, have argued that because many ELs and their families are unfamiliar with the U.S. educational system, and because ELs tend to score poorly on language-heavy exams, ELs are often placed in classes that are remedial or do not prepare them for college. As a result many of them fall further and further behind native English-speaking peers with the same academic capacity. All in all, this snapshot begins to reflect why school districts and states feel enormous pressure and often lack of readiness to provide viable schooling for their student bodies (García, Jensen, & Scribner, 2009).

In particular, there appear to be two overarching challenges to appropriately measuring the academic achievement of ELs: (a) proper exposure to challenging content in school and (b) proper evaluations and assessments that minimize their English language limitations and cultural misunderstandings while being able to still effectively measure their knowledge and skills in subjects such as mathematics and science.

Policy overview. The Civil Rights Act of 1964 advanced the federal commitment to equity in education and in 1974 *Lau v. Nichols* spelled out the

UNCORRECTED PROOFS © AMERICAN PSYCHOLOGICAL ASSOCIATION

APA-HTA_V3-12-0603-018.indd 371 12/10/12 1:30 PM

educational rights of language minority students. Up until the 1994 reauthorization of the federal ESEA, however, most ELs were exempted from most state and local standardized achievement testing regimens and, with little accompanying accountability oversight, were often schooled separately from their native English-speaking peers. This exclusion changed in 1994 and again in 2001 when the NCLB reauthorization was passed, and states and schools were held accountable for ELs in such a way that teachers were expected to teach, and ELs were expected to learn, the same academic content as their native English speakers. Once this change occurred, researchers and practitioners began to investigate how to make challenging content and assessments accessible for this population.

To design and build accessible achievement assessments several interwoven steps are essential. This is particularly the case when the assessment systems are constructed to measure the same concepts and skills of all test takers, including but not limited to ELs and SwDs. The rest of the chapter outlines and discusses some of the primary issues and solutions that have been found to be effective to date. These and other considerations are discussed in more detail in Kopriva (2008).

BUILDING ACCESSIBLE SYSTEMS: SETTING THE STAGE

Before test construction begins, it is important to put into place procedures associated with participation in test development and methods to ensure that items and forms are accessible.

Participation in Test Development

Adequate participation of EL and SwD experts as well as adequate representation of EL and SwD students should to be built into the development process. Typically, experts with substantive knowledge of these populations have been used primarily in bias reviews, where the charge has been narrow. They have not been included in the planning, item development, and decision-making processes to the same extent that mainstream teachers and content experts have been in recent years. This participation includes involvement throughout the design, construction, and technical phases of development (for a general discussion of test development, see Volume 1, Chapter 9, this handbook). Tasks in which it would be appropriate for them to actively participate can be found in Exhibit 18.1.

Experts who bring the most to the test development process have a deep understanding of content standards, experience with adapting academic teaching environments for these students, and knowledge of their students' strengths and challenges. Examples of relevant expertise of EL experts can be found in Exhibit 18.2.

Exhibit 18.1
Expert Participation

- Designing the comprehensive testing system
- Developing test specifications
- Writing and reviewing content items and rubrics that are appropriate for the students with disability and English learner populations
- Providing training to other item writers and developers
- Trying out items in classes
- Evaluating forms for coverage and accessibility
- Making decisions about inclusion or exclusion of items, all testing materials, and administration and response options based on data from pilots, field tests, and other technical data collections
- Scoring, reporting, and making decisions about test use for accountability and program evaluation

Exhibit 18.2
Types of Expertise for English Learner Experts

- Educators from classrooms in which students are learning English as well as grade-level academic content
- Educators from mainstream academic classrooms in which English learners are placed after they have reached a certain level of English proficiency
- Educators working with students who are newly arrived to the United States
- Educators working in classrooms in which the students' primary language (also known as their first language or L1) is the language of instruction or in bilingual (L1 and English) classrooms
- Educators with urban experience and educators with rural experience
- Educators working with migrant students
- Educators who come from the primary language and cultural backgrounds of the students they teach

UNCORRECTED PROOFS © AMERICAN PSYCHOLOGICAL ASSOCIATION

APA-HTA_V3-12-0603-018.indd 372 12/10/12 1:30 PM

Just as the diverse perspectives of multiple experts should be included during test development, a full range of SwDs and ELs should be involved in all item and test data collections. It is well known that ELs respond differently based on their proficiency levels and adequate accommodations, and so participating students should range from new arrivals through former English language learners that have successfully transitioned. The same is true for SwDs, whose diverse set of challenges make this broad category extremely heterogeneous. To ensure validity of inferences across all tested students, it will be important to determine that all subgroups are responding in similar fashion. As such, enough ELs and SwDs from preidentified strata should be included during piloting to be able to analyze the data by these subgroups as well as the mainstream population. Sireci and Wells (2010) and DePascale (2010a), among others, recommend that the analyses should control for academic ability, and they have demonstrated several ways this might be accomplished.

Building in Procedures to Ensure an Accessible Product

Kopriva (2008) argued that ensuring access is not just a post hoc project. Rather, in addition to including SwD and EL experts and students in range of development, it is important to explicitly consider during planning if general items and forms are accessible, and, if so, for whom. When accommodations will be used, have the proper accommodations been selected and for which EL and SwD student profiles, and is there an oversight mechanism in place to ensure that each student is receiving what they need during the test administration? Are their translations of any forms, are they of high quality, which ELs will they benefit, precisely, and who is still not accommodated adequately? For which SwDs, precisely, are modifications or alternate assessments being considered, and are the plans adequate to satisfy their accessibility to the academic content? Procedures, such as conducting bias reviews and analyzing differential functioning of some items in some subgroups, are seen as ways to address accessibility but are not sufficient by themselves. Finally, to ensure that all questions such as these are adequately addressed, Kopriva maintained that test publishers and consumers should develop a systematic system for checking that the needs of all students are properly considered in test development. This system is briefly outlined in the last section of the chapter.

PRINCIPLED ITEM AND FORM CONSTRUCTION

For many students (e.g., many ELs and some SwDs with literacy, language, or attention or other disabilities), how items in standardized testing systems are typically presented and communicated to the students represent barriers to either accessing what the item is asking or barriers to how the student can show what they know. In these cases, accessible forms with item adaptations need to be created to minimize the barriers and measure intended content at specified cognitive complexity levels.

For most SwDs and all ELs, item adaptations in standardized content testing systems are purposefully designed to measure the same content and cognitive skills as the general test that is given to a majority of the student population. In these cases, if properly constructed, adapted forms and formats and general test forms are intended to yield the same score inferences about the same knowledge and skills. In some cases, however, some SwDs are assessed in large-scale statewide content assessment systems (and some other systems) using modifications and alternate assessment forms and formats that are known to result in different score inferences. The decisions to measure content with modifications or alternate assessments are driven by the nature of the students' disabilities. Both adaptations built to be interchangeable with general test forms and those considered to not be interchangeable are briefly discussed.

Item and Form Adaptations Built to Be Interchangeable

Accessible forms with item adaptations measure the same content, at the same cognitive complexity, as the items used in the general test, and provide, as necessary, alternative ways for students with particular needs to meaningfully respond. Form and item adaptations may include braille or

UNCORRECTED PROOFS © AMERICAN PSYCHOLOGICAL ASSOCIATION

APA-HTA_V3-12-0603-018.indd 373 12/10/12 1:30 PM

large print, translations into languages other than English, plain-language edits in English, and use of such supports as visual aids or access to such tools as manipulatives or picture glossaries. Harnessing computer capabilities increases how meaning might be successfully conveyed, for instance, through animations and interactive aspects, and this methodology also allows for greater flexibility in how students can respond, for instance, by demonstrating their skills, assembling, or modeling (e.g., Kopriva, Gabel, & Cameron, 2011).

Form and item adaptations designed to measure the same content and cognitive complexity and lead to the same score inferences as the original or base items and forms share certain key development processes, regardless of the nature of the adaptations. First, using a model such as Evidence Centered Design (ECD; Mislevy, 1996), a clear understanding of what the intended inferences are at the item level is essential. Note that explanations of the target content and complexity at the item level are at a finer grain size than is typically required in general tests but are necessary if student scores on the adapted forms are going to be considered interchangeable. Second, particular barriers and then particular item adaptation elements intended to ameliorate or minimize each barrier need to be identified. Third, using techniques that have been found to be successful item adaptations can be designed and built to address one or more particular barrier purpose while still measuring the same content and processes, at the same levels of cognitive complexity. Contextual concerns, formatting, layout of text and nontext elements, attention to language and linguistic structural factors, and continuing adherence to meaning in the base item are always considered.

Adapted forms often have a similar look and feel as the general forms, albeit in a different language or in large print. With enough documentation, however, it can be argued that forms such as portfolio systems (Barton & Winter, 2010) or computer-interactive animated forms and tasks (see Kopriva, Gabel, et al., 2011) may be used to yield similar score inferences as general forms (on paper or on computer) with multiple-choice and constructed-response items. It remains an ongoing discussion in education assessment as to what kinds of evidence and documentation are needed to successfully make these cross-forms and format claims. For instance, how comparability issues are resolved can influence these adaptations that do not adhere to the given structure of the general test. Readers are encouraged to read the last section in this chapter for a summary of some of these issues as well as Volume 1, Chapter 4, this handbook.

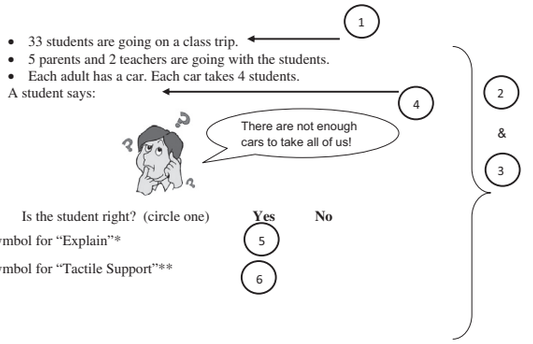
Although readers are directed to other resources for details on how to properly design adapted items (e.g., see Kettler, Elliott, & Beddow, 2009; Kopriva, 2008; Thurlow, Thompson, & Lazarus, 2006), one example of a plain-language edited item in English with formatting and visual supports, and response adaptations, can be seen in Figure 18.1). This item and its base (Figure 18.2) were used in a randomized study of elementary students (Kopriva & Mislevy, 2005) and are discussed in Kopriva (2008, Chapter 5). Independent ratings found that both items measured the same content and targeted cognitive complexity. The study found that the adapted item measured the targeted mathematics ability significantly better than the base item for many ELs and for struggling native English-speaking readers with no individualized education programs (IEPs), whereas both items similarly measured the targeted ability for more robust native English-speaking readers. These findings suggest that adaptations such as what are shown in Figure 18.1 can be effective in measuring the intended inferences for students that present certain barriers to communication similar to what are illustrated in Figure 18.2. Explanations of some of these adaptations are shown in Figure 18.1.

Item and Form Adaptations Built to Lead to Different Score Inferences

For SwDs, the student's IEP team makes the decision as to how the student will participate in large-scale academic assessments systems. For a relatively small number of these students, these recommendations involve the use of modifications or one of two types of alternate assessment forms.

UNCORRECTED PROOFS © AMERICAN PSYCHOLOGICAL ASSOCIATION

APA-HTA_V3-12-0603-018.indd 374 12/10/12 1:30 PM



*This symbol is introduced before the test and is common across all tests this state uses. It means students need to provide an answer and they can do so using words, algorithms, pictures, or other diagrams.

**This symbol is introduced before the test and is common across all tests this state uses. It means that there is an available tool set students can tactilely manipulate to help them solve the problem.

1. Information that is not needed to set the context for the problem has been eliminated, reducing the amount of text.
2. Plain language principles have been applied to the item to reduce the semantic and syntactic complexity of the item. The sentences are shorter and straightforward, using present tense and active voice and reducing the use of prepositional phrases and dependent clauses. A visual is used to illustrate the item. Note that numerals have been used consistently throughout. The translation between a verbal and symbolic representation of a number was considered construct-irrelevant mathematics.
3. The formatting has been arranged to provide maximum access to the problem requirements. Each complete piece of information is presented separately, since, for this item, selecting the appropriate information from among relevant and irrelevant pieces of information was not part of the measurement target. The question is clearly separated from the rest of the text, and the two-stage character of the item, answering the question and explaining the response, is evident.
4. While both the base and the variation assume students are familiar with class trips, which may not be the case in all schools, potential cultural schooling bias has been reduced in the variation by having a student's statement the focus of the question. In some cultures, children are not used to questioning teacher judgments and decisions .
5. Students are given options for how they represent their response.
6. Students are allowed to use manipulative tools to help them represent and solve the problem. The targeted content knowledge and skills do not preclude allowing various methods of representation or solution. The manipulatives provide students who are ELs a way to represent the text that may help them understand the problem situation.

FIGURE 18.1. Adapted item.

At Jefferson Midlands Middle School, the sixth grade students and their teacher are planning a field trip to the state capital at the end of the year. In the morning they will visit the state legislature, and in the afternoon they will go to the zoo.
There are 33 students in sixth grade. Five parents and two teachers will be coming with the students on the trip. Each of the adults has a car that can hold four students. One of the teachers says: "There are not enough cars to take all of us!" Do you agree with the teacher? Explain your answer.

FIGURE 18.2. Grade 4 Mathematics item (base).

Modifications. Like the presentation adaptations, modifications provide alternatives to the standardized way test forms are presented to some SwDs to allow some students to better demonstrate their knowledge and skills in learning and testing situations. These adaptations, however, change the testing situation in a way that changes the construct being measured, and hence because of purpose and use, they are generally defined as modifications versus accommodations (Thurlow et al., 2006). Modifications can be made with respect to test modality, complexity, space, time, language, and possibly other aspects (Poteet, 1990). This means that some modifications are form related, whereas some involve other accommodations, which are discussed in the next section. Form-related modifications may involve substituting some of the general test items with modified items that are less cognitively complex, using fewer option choices in multiple-choice questions, or scaffolding constructed-response items that may change the nature of the targeted construct if the constructs involves skills associated with how to approach and conceptualize the problem-solving process.

Specifications for modifications are considered part of NCLB's federal peer review guidance associated with statewide content K–12 assessments used for accountability purposes. Modification constraints may be identified by test publishers, or by users of district or other tests, such as the SAT (formerly known as the Scholastic Aptitude Test) or

UNCORRECTED PROOFS © AMERICAN PSYCHOLOGICAL ASSOCIATION

APA-HTA_V3-12-0603-018.indd 375 12/10/12 1:30 PM

Graduate Record Examinations (GREs). Consequences are varied and sometimes convoluted, but some consequences include "flagged" test scores—that is, scores that are not allowed for accountability purposes—and scores that count for some purposes but not others (Thurlow et al., 2006).

Alternate assessments. These assessments are intended to facilitate evaluations of academic content knowledge and skills for students with significant disabilities. This type of content assessment was specifically defined within the original NCLB legislation, additional regulations, and nonregulatory guidance; tends to focus on particular purposes and uses; and addresses certain criteria. These assessments are not assumed to be interchangeable with general test forms, but the scores from alternate assessments are used as evidence of statewide accountability for federal purposes. Two forms of alternate assessments currently exist. Albers (2011) recently developed an alternate assessment form of the ACCESS for ELLs to measure the English language proficiency of ELs who also have significant cognitive disabilities.

Ideally, alternate assessments should provide instructional utility, guiding the development of future instructional goals and learning. Thus, alternate assessments should meet needs for both *required* information (i.e., for accountability) and *desirable* information (i.e., for instructional utility). The assessments must meet the same standards of high technical quality—validity, reliability, accessibility, objectivity, and consistency—expected of other educational tests. In addition, alternate assessments of academic content must have an explicit structure, guidelines for determining which students may participate, clearly defined scoring criteria and procedures, and a report format that communicates student performance in terms of academic achievement standards.

Alternate assessments for students with the most significant cognitive disabilities. Students with the most significant cognitive disabilities are individuals who (a) have disabilities

> within one or more of the existing categories of disability under the IDEA (e.g., autism, multiple disabilities, traumatic brain injury, etc.), and (b) whose cognitive impairments may prevent them from attaining grade-level achievement standards, even with the very best instruction. (U.S. Department of Education, 2005, p. 23)

This type of alternate assessment is based on *alternate* achievement standards. These standards are required to be aligned with grade-level content standards, but they are allowed to be reduced in depth, breadth, and complexity. The U.S. Department of Education allows up to 1% of a school district's total number of students to be rated as "proficient" or "advanced" using alternate assessments that are based on alternate achievement standards.

Alternate assessments for students with other significant disabilities. Additional regulations were established in April 2007 that allow states to report proficient or advanced scores for up to 2% of the total student population using alternate assessments based on *modified* achievement standards. Alternate assessments based on modified achievement standards are directed toward a small group of SwDs who have been determined to be capable of making significant academic progress, but who nonetheless may have significant difficulties in reaching grade-level achievement. In contrast to the *alternate* achievement standards on which students with the most significant cognitive disabilities may be assessed, modified achievement standards are not based on a restricted range of grade-level content. They are based on the same range of grade-level content as the general achievement standards, although the expectations for mastering the grade-level content standards may be less rigorous. An alternate assessment based on modified achievement standards, for example, may include less difficult items based on the same content as the general assessment, include fewer distractors on multiple-choice questions (e.g., three response choices rather than four), or have shorter reading passages than the general assessment.

ADDITIONAL TEST ACCOMMODATIONS

Test accommodations sometimes refer to testing condition adaptations that fall outside of what is

UNCORRECTED PROOFS © AMERICAN PSYCHOLOGICAL ASSOCIATION

APA-HTA_V3-12-0603-018.indd 376 12/10/12 1:30 PM

presented to students. Like the adaptations in items and forms, these changes are based on minimizing particular barriers and are used most often for ELs or SwDs. In the language of current educational testing, the term *accommodations* refers to changes in conditions that do not alter the construct being measured; *modifications* refer to changes in conditions that do alter the construct. For ELs, in addition to adaptations to forms and items, accommodations include tools, administration, and response accommodations (Abedi, 2007). For SwDs, additional assessment accommodations usually fall under one of the following categories: equipment and materials, scheduling and timing, setting, linguistic, and response accommodations (Christensen, Lazarus, Crone, & Thurlow, 2008). It is well known that the permissibility of specific accommodations varies across content area and state or other users. These differences have led to a great deal of confusion particularly when cross-educational agency comparisons are made (Fields, 2008).

The following sections outline some of the most relevant accommodations for both SwDs and ELs. Readers are directed to Kopriva (2008) for a fuller explanation of and research base for EL accommodations, and to Thurlow et al. (2006) for more details and additional resources about accommodations for SwDs.

Tools

For SwDs, common equipment and materials accommodations alter the test setting to include certain types of tools and assistive devices, including magnification equipment, amplification equipment, templates, and lighting or acoustics. For ELs, tools often include bilingual, English and picture glossaries, and sometimes manipulatives and other content-relevant materials used by students to demonstrate what they know without using much language.

Administration Accommodations

Primary administration accommodations for ELs involve oral English or oral administration of the assessment in their home academic language. Secondary administration accommodations for this population are specified to facilitate the oral administrations or response demonstrations or to deal with extended time requirements, anxiety, or fatigue. They generally include extra time, small group or individual administration, and more frequent breaks. For SwDs, administration accommodations include signing and interpreting directions and reading questions aloud. They might also include scheduling and timing accommodations such as change of time or scheduling of a test, incorporating breaks, testing at a time that is beneficial to the student, and allowing extended time. Administration accommodations might also involve setting—for instance, changing the test location or environment (including individual or small-group administration or administration in a separate room or carrel) and changing the proximity of the student's seat to the test administrator.

Response Accommodations

Response accommodations, as they are defined most often, change the standard conditions around how students can respond to the items presented to them, including the parts of the items presented to them that frame the response environments. Like administration accommodations, these post hoc adaptations do not change the response options or forms of response themselves—any substantive variations that alter the kinds of responses students can reply to are item adaptations. For SwDs, examples of post hoc response accommodations include using a Brailler, writing in test booklets, and using a computer or machine to communicate what the students know (including not only disability-specific technology such as recording puffing or visual cues and then translating these data into a form that can be scored but also using a tape recorder or voice recognition that records the students' audio responses). Communicating responses to a proctor or scribe and allowing this person to bubble, complete, or write the response is another accommodation used for some SwDs.

For ELs, response accommodations have typically involved students responding orally or in text using their home language or code-switching (using both English and their home language). Although these methods seem to be effective for constructed-response

UNCORRECTED PROOFS © AMERICAN PSYCHOLOGICAL ASSOCIATION

APA-HTA_V3-12-0603-018.indd 377 12/10/12 1:30 PM

items (e.g., see Kopriva & Mislevy, 2005), they do not affect multiple-choice or other close-ended items that make up the vast majority of standardized tests. Item variations that use better editing, plain-language text, and visual supports help students with higher English proficiency respond meaningfully to these types of questions, but these methods are often not enough for students with lower English proficiency. Although the multiple-choice questions do not require any additional language to respond, the language of the options is often problematic. For students with little English and for those with little literacy in their home language or first language (L1; in cases in which the test is in L1), correct response to these questions hovers around the guessing level, making this type of item a bad fit for these students (Emick, Wiley, & Kopriva, 2007). An adequate accommodation would be to allow these students to communicate by demonstrating or modeling their knowledge and skills rather than using only English language, but this approach is usually not feasible in high-volume testing. Recent large-scale prototypes of computer-interactive test questions that allow these students to demonstrate, assemble, and model what they know have been found to be very effective (Kopriva & Carr, 2009). Efforts are under way to integrate these advances into large-scale summative and formative testing systems.

ASSIGNMENT OF TEST, FORMS, AND ACCOMMODATION OPTIONS

Even as large-scale content tests may be developed and accommodated to specifically address the needs of ELs and SwDs, if there is no technically rigorous mechanism in place to get the specific methods to the specific students who need them, it is argued that these efforts have little effect. Several researchers who investigate accommodation effectiveness for these populations point out that consistent and appropriate accommodations decision making is critical to the validity of standardized academic testing programs and to the ability to properly use scores to compare student performance across states and districts (e.g., Fuchs, Fuchs, Eaton, Hamlett, Binkley, & Crouch, 2000; Hollenbeck, Tindal, & Almond, 1998; Kopriva, 2008). At the individual level when accommodations decisions are not appropriate to meet the needs of individual students, test results misrepresent their knowledge and skills (Hipolito-Delgado & Kopriva, 2006). At the aggregate level, when accommodations decisions are inconsistent from classroom to classroom or district to district, comparisons across classrooms, districts, and states may be unfair and meaningless (Abedi, 2007; Fields, 2008; Solomon, Jerry, & Lutkus, 2001).

Current guidelines for selecting large-scale and classroom-based accommodations for content testing of SwDs primarily stems from authorizations of federal legislation in IDEA. Regulations or instructions for assigning accommodations to individual ELs, on the other hand, are generally policy based, most often at the state level. The practice for assigning large-scale accommodations for SwDs typically focuses on the role of the IEP. In addition to developing and evaluating each student's learning goals and instructional plans, the IEP addresses the proper test accommodations appropriate for each student at both the classroom and standardized testing levels. Current practices typically used to assign large-scale test accommodations to individual ELs reflect that decisions generally are made by a single person (commonly the student's teacher or the school EL specialist), although some education agencies are beginning to use teams.

In both situations, guidelines tend to offer broad parameters rather than specific guidance for those who must make accommodations decisions. Both individual teachers and teams making accommodations decisions attempt to work within the policies given to them by the federal, state, or local education agency, but these policies generally do not contain specific recommendations for how to address the needs of specific students. Koran, Kopriva, Emick, Monroe, & Garavaglia (2006) found that teacher recommendations, unfortunately, were not statistically different from random assignment of large-scale content testing accommodations to EL students. In the past few years, there have been efforts to tighten the criteria for accommodating SwDs and ELs (e.g., Fields, 2008), but large inconsistencies remain at all levels of schooling.

UNCORRECTED PROOFS © AMERICAN PSYCHOLOGICAL ASSOCIATION

APA-HTA_V3-12-0603-018.indd 378 12/10/12 1:30 PM

Research over the past 10 years has continued to confirm that one cannot validly assign accommodations to groups of students based on some broad classification or status (Sireci, Li, & Scarpati, 2003). How then should educators intelligently and reasonably make decisions about accommodations for particular SwDs and ELs when competing tensions of time and accountability are combined with the complexity of needs associated with the heterogeneous populations?

Emerging work suggests that systematic methods of assignment may work better than relying on current policy approaches to assign accommodations for both SwDs and ELs (Fuchs, Fuchs, Eaton, Hamlett, Binkley, et al., 2000; Helwig & Tindal, 2003; Kopriva, Emick, et al., 2007; Russell, 2010; Weston, 2003). Furthermore, researchers present evidence that using systematic methods to match the particular needs and strengths of individual students to specific accommodations may increase validity and be superior to using educator-directed decision making alone.

Elliott and others (e.g., Elliott, Kratochwill, & Gilbertson-Schulte, 1999; Roach & Elliott, 2006) have continued to provide guidance to IEP teams about how to wisely assign large-scale accommodations for SwDs. These researchers have identified key information and student needs that teams should know, critical access skills that are particularly salient for this population, and process factors that influence accommodation decision making. The *Assessment Accommodations Guide* (Elliot et al., 1999) and associated guidance direct IEP team members through the accommodation selection, implementation planning, and documentation processes. The authors encourage members to link any of the 16 key access skills they have identified as being problematic for an individual student to one or more accommodations that specifically minimize interference between conditions and measurement of target skills. These skills represent elements of typical large-scale standardized testing conditions that could pose a problem for SwDs.

The work of Fuchs and colleagues (Fuchs, Fuchs, Eaton, & Hamlett, 2005; Fuchs, Fuchs, Eaton, Hamlett, Binkley, et al., 2000) that is discussed briefly in the next section provides a good example of an empirically grounded systematic method for matching SwDs to particular accommodations based on specific needs. These researchers have found that the test scores of SwDs who receive appropriate accommodations reflect more accurately what others think these students know and that their method is far superior to other methods of accommodation matching. This method is time intensive, however, as determinations are made individually using a trial-and-error process.

More precise, systematic, guidance to identifying needs of SwDs and then recommending some classes of large-scale accommodations for content tests was completed in the past few years by a consortia of states (Christensen, Thurlow, & Wang, 2009). To date, however, there is little research to support that it is consistently better than state-level guidance manuals that several researchers, including Thurlow and colleagues (e.g., Thurlow, Lazarus, Thompson, & Robey, 2002; Thurlow, Moen, & Wiley, 2004; Thurlow et al., 2006) and Rivera and Collum (2006) have argued lead to notoriously inconsistent assignments over locales and students with similar profiles. A recent program that guides IEP teams through a long series of student needs and prior accommodations questions and then leads to suggestions for particular accommodations may be more successful (South Carolina Department of Education, 2010).

Recently, Abedi (2007) and Rivera and Collum (2006) have introduced a hierarchy of choices to the large-scale accommodations for ELs. As discussed in the previous section, the researchers divided relevant accommodations into primary and secondary. The primary accommodations refer to language adaptations, whereas the secondary accommodations refer to conditions that can facilitate or at least not discourage the ability of students to receive the primary accommodations. Rivera and others in her center have compiled guidance to encourage teachers to properly choose accommodations based on guidelines around these primary and secondary accommodation sets, but research about the effectiveness of this advice is as yet unpublished.

To date, only one systematic accommodation matching system for ELs has been published. *STELLA*, the *Selection Taxonomy for English*

UNCORRECTED PROOFS © AMERICAN PSYCHOLOGICAL ASSOCIATION

APA-HTA_V3-12-0603-018.indd 379 12/10/12 1:30 PM

Language Learner Accommodations, is a newly developed informant system designed to assign individual accommodations for K–12 ELs (e.g., Carr, 2009). It identifies critical variables, collects data, combines the data with standard information regarding how accommodations perform, and then uses a standard series of computerized algorithms. These algorithms have been successfully built, revised, and vetted by experts (Kopriva & Hedgspeth, 2005) and by a team of state specialists (Carr, 2008). The system is designed to utilize the latest information about students that appear to be the most relevant for making accommodation decisions about this population, and it is designed to be customized to accommodate the policies of different states or districts. One of two validation studies found that this system seems to be producing decisions for individual students that better match the data than teacher methods do (Koran et al., 2006), whereas the second study found that ELs who received proper accommodations scored significantly higher than ELs who received incorrect or no accommodations (Kopriva, Emick, et al., 2007).

Guidance manuals that leave the decisions to teachers or IEP teams do not seem to be sufficient to ensure ongoing consistency across locales. The STELLA computer-based matching method looks promising for ELs, whereas trial-and-error methods, or taxonomies of pointed questions for guiding decision makers, may be more relevant for SwDs. Whatever effective processes are used, Solano-Flores and Trumbull (2008) have argued that they must be coupled with consistent implementation procedures and systematic oversight or else the benefits of appropriate matching may be lost. In one hint of how this might be accomplished, Russell (2010) reported that he is currently working with a test publisher to link recommended accommodations directly with computer-based tests for SwDs in such a way that students would receive some of their accommodations electronically as they take their content assessments. Going forward, it will be important to continue to focus on refining consistent data-collection methodologies that isolate the most relevant information for decision making and to continue to attend to the algorithms that are used to convert and combine data and to the decision-making rules to ensure that they sensitively yield the most salient accommodations for the students who need them.

TECHNICAL CONSIDERATIONS

This section focuses on three interrelated issues that need to be considered to defend the scores from content tests that include variations that address the testing needs of SwDs and ELs: (a) defining and building content assessment systems with proper adaptations, (b) conducting interpretable research, and (c) constructing empirically based comparability arguments to support when scores should and should not be considered interchangeable.

Defining Content Assessment Adaptions

Kopriva (2008, Chapter 12) described an adapted evidence-centered design model and procedures for test developers to use when building their assessment systems to include variations for these populations. The approach is designed to identify which item and form, tools, administration, and response adaptations to make and use in assessment systems when interchangeable score inferences are intended. Specifically, beginning with (a) identifying intended inferences, the approach recommends methods for (b) identifying the assessment barriers for various profiles of ELs and SwDs, (c) identifying the variations to address the specific barriers for specific profiles, (d) constructing the tests and additional accommodation options to include the variations, and then (e) employing oversight procedures to ensure that all intended adaptations are included, that appropriate students receive the proper adaptations, and that the proper analyses are completed to support the common inferential claims. As noted, attention should be paid at the item as well as form and post hoc accommodation levels to support construct validity and comparability arguments for students who take the large-scale tests under nonstandard conditions. Often, in our rush to build content assessment systems, the design procedures step is shortchanged, putting the framework for the entire assessment system at risk. Going forward, others as well as Kopriva (e.g., Barton & Winter, 2010; DePascale, 2009; Winter & Gong, 2009) have asked for thoughtful and organized a priori designs

UNCORRECTED PROOFS © AMERICAN PSYCHOLOGICAL ASSOCIATION

APA-HTA_V3-12-0603-018.indd 380 12/10/12 1:30 PM

to determine which adaptations are to be integrated into or recommended for content testing systems, and how these systematic specifications, implemented properly, might interact with and help defend notions of comparability.

Conducting Interpretable Research

Several authors, including Abedi (2007), Rivera and Collum (2006), Thurlow et al. (2006), and Tindal and Fuchs (2000), have described the types of item and form, tools, and post hoc accommodations that seem to be useful for SwDs and ELs who have particular profiles of needs and strengths. Yet, research findings that underpin effective links between student profiles and accommodations are often mixed (Kopriva & Lara, 2009), particularly for ELs. To some extent, the fault lies in studies conducted without the proper robust research controls as consumers and test developers rushed to implement accessible agendas quickly. Furthermore, lack of research funding, the heterogeneity of the SwD and EL populations, and the small numbers of many students who fit certain profiles at specific grades or content areas makes research difficult. But methodological flaws, funding constraints, small populations in some cases, and small sample sizes in many of the studies are only part of the story.

Students with disabilities. Tindal and Fuchs (2000) asserted that for accommodation effectiveness to be considered defensible, these accommodations should be based on individual need. The accommodations should benefit only or mainly the students who need the change and not other students. To address the first part of this challenge for SwDs, Fuchs and colleagues (Fuchs, Fuchs, Eaton, Hamlett, & Karns, 2000) set out to conduct a series of single-subject investigations in which students with certain profiles were given no accommodations and then one or more in sequence, checking at each point to see when the students responded in a manner that the researchers thought was closer to what the student actually knew. If a boost in response rate was evident with one set of accommodations versus another, the researchers concluded that this set was the proper adaptations for this type of student. Over time, many of these SwD profile and accommodation choices were assembled and Fuchs et al. (2005) published the *Dynamic Assessment of Test Accommodations*. These profiles, however, do not nearly cover the range of SwDs.

What has substantially lagged for this population are focal and control group experimental investigations (groups that receive and do not receive accommodations) to address the second point made by Tindal and Fuchs (2000). This type of research is long overdue. Most of the group accommodation studies to date are post hoc and outside well-designed and systematic research agendas, both of which makes the findings less tenable. Although small populations and sample sizes of low-incidence groups make implementing the research challenging, there are only a few investigations of this type with high-incidence populations. It is suggested that experimental studies for low-incidence populations be conducted over years or over sites, using the same experimental design and comparable controls. Even then, most aggregate samples will be small in nature.

Albeit with a single-subject research design, Fuchs, Fuchs, Eaton, Hamlett, and Karns (2000) have laid the groundwork for research on effective accommodations for SwDs. It seems reasonable that with such a heterogeneous population, streams of group experimental research need to be organized to take advantage of the work that has been done and to build a directory of findings that can defend the kinds of test adaptations consumers and advocacy groups are expecting for students with particular needs.

English learners. The accommodations research for ELs has followed a different path. Consistently, meta-analyses (Kieffer, Lesaux, Rivera, & Francis, 2009; Pennock-Roman & Rivera, 2006; Sireci, Li, & Scarpati, 2003) find many questionably designed studies and only a small number of experimental investigations. Results are mixed, even for well-controlled studies. Why is this? Kieffer et al. (2009) have argued that perhaps accurate content inferences from large-scale testing (even with accommodations) are not possible for students with lower language proficiency. Others, however, have argued against this (Grant & Cook, 2010; Kopriva, Lundberg, & Boals, 2011). Instead, they

UNCORRECTED PROOFS © AMERICAN PSYCHOLOGICAL ASSOCIATION

APA-HTA_V3-12-0603-018.indd 381 12/10/12 1:30 PM

have asserted that much of the confusion is a casualty of how the groups are defined and how the studies are designed. Most centrally, almost all the studies to date have studied ELs as a monolithic group even though researchers are aware of the diversity of their needs and strengths (Abedi, 2011; Kopriva, 2008; Solano-Flores, 2010). Not surprisingly, focal accommodations that may be effective for one subgroup of ELs is often not useful for another. Several researchers (e.g., Abedi, 2007; Emick & Kopriva, 2006; Emick et al., 2007; Kopriva, Emick, et al., 2007) have argued that level of English language proficiency, at the very least, is a group criterion—students with low English proficiency often need different accommodations than those with higher proficiency. Other characteristics appear to be important as well, such as literacy in their home academic language and how they have been schooled to date (Carr, 2009). Kopriva, Emick, Hildago-Delgado, and Cameron (2007) illustrated that English language proficiency and L1 literacy were salient factors in choosing proper accommodation sets.

Furthermore, Kopriva, Cameron, and Gabel (2010) found that providing adequate nontext language rollovers, some L1, and a broader set of response avenues were effective in measuring the science of ELs with the lowest English proficiency, to such an extent that they scored on par with their native English-speaking peers. Examples of nontext language rollovers include static or animated visuals or halo-highlighting of relevant areas on the screen. To address Tindal and Fuch's (2000) second point, it is interesting to note that the English-speaking peers in this study did not score significantly differently on adapted items than they did on the general test form, whereas EL scores were substantially higher on the variation as compared with the general test. This study is significant because it suggests that large-scale testing *can* be properly accommodated for even students with very little English or literacy skills.

All in all, only a few studies with proper grouping have been completed. Until there is a critical mass, it will be difficult to make definitive judgments about the usefulness of specific accommodations or accommodation sets for ELs with particular profiles.

Comparability Evidence That Supports Decisions About Scores

Advances in cognitive learning theory in the 1990s led to the identification of an expanded set of measurement approaches that seemed to be promising for use in large-scale content assessment. The focus was on comparability of responses *within* approaches—for instance, when rubrics allowed for various ways for students to demonstrate their content knowledge and skills at, say, a Level 3 out of 4 possible points. Over the past 15 years, federal legislation mandating inclusion challenged the status quo that required all students to take tests under standard conditions. This required considering when scores from tests taken under various conditions by SwDs and ELs might be considered interchangeable. Mislevy (1996) argued that the traditional argument for common inferences was made on procedural grounds, leading to the requirement for common products and testing conditions. It is the common inferences, however, that test developers are interested in holding constant, not the procedures per se. As such, Mislevy and others (e.g., see Mislevy, Steinberg, & Almond, 2003) suggested that this conceptual argument should be built on providing adequate evidence about the knowledge and skills of interest, necessary observations, properties of tasks or items designed to elicit the observations, and assessment situations in which students interact with assessment requests. This approach suggests that data may be collected under alternate conditions, as long as there is proper documentation and evidence.

A number of issues relating to validity and comparability are discussed in Kopriva (2008, Chapter 12). Readers are encouraged to review this chapter for more detailed information relevant to making decisions about comparability when different testing conditions are used for different students or when variations in forms are considered. Additionally, related chapters include Volume 1, Chapters 4 and 17, this handbook, and Chapter 17, this volume. What follows is a brief summary of some of the primary points associated with comparability of scores in K–12 academic content testing for ELs and SwDs.

For the purposes of topics discussed in this chapter, comparability of score inferences suggests that

UNCORRECTED PROOFS © AMERICAN PSYCHOLOGICAL ASSOCIATION

APA-HTA_V3-12-0603-018.indd 382 12/10/12 1:30 PM

the meaning of the scores is the same, whether students take form A or form B. Comparability includes two steps: First, development methods and empirical evidence need to demonstrate that the forms are measuring equivalent knowledge and skills in the content domain of interest. Second, assuming content equivalence across forms, statistical methods place scores from the forms on a common scale so that comparisons can be made across forms. The *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) specifies that comparability is easiest to establish when procedures for test development and implementation are identical. In cases in which formats or other conditions vary, other evidence may be required. In all cases, the requisite degree of comparability is defined in terms of context and purpose for which the tests (or forms) are being used. For example, to compare the performance of individual students over time or with other students, a more precise definition of comparability would most likely be required than when equivalence is determined at the prespecified achievement standards level (e.g., basic, proficient, and advanced) with school-level data (i.e., how adequate yearly progress is reported under the NCLB legislation).

In their discussion of steps for achieving comparability when accommodations or nonstandard forms are used by some students, Haertel and Wiley (2003) focused on the necessity of determining target equivalence before any consideration of statistical equivalence is considered. Haertel (2003) differentiated between comparability of objectives for norm-referenced and criterion-referenced tests (like the standards-based assessments being used today) and the implications for these tests when standard and nonstandard forms are considered. He also specified comparability of test administrations under student-to-student comparisons, student-to-achievement level standards comparisons, and student-to-quantifiable criterion measures. Although he suggested that judgments may be the primary avenue when student score inferences are desired at the level of achievement standards, he did not explain how comparability might be accomplished across nonidentical forms when more precision within achievement levels is desired. Winter and Rabinowitz (2006) defined two conditions, both of which they believe are necessary to evaluate comparability. The first of their two conditions is construct consistency: At the targeted level of comparison (i.e., to other students or to content standards), do the forms measure the same content information? The second condition is that of score consistency: At the appropriate level of comparison, do the same scores or same performance levels, across forms, reflect the same level of abilities? They emphasized that adequate evidence is essential to document the equivalence at each level.

Winter and Rabinowitz (2006) argued that only after an adequate level of content equivalence has been established, should score equivalence methodologies be implemented. Mislevy (1993) differentiated three levels of linking academic forms—equating, calibration, and social moderation. Feuer, Holland, Green, Bertenthal, and Hemphill (1999) extended these methods to include equating, calibration, projection, and moderation. In both taxonomies, the methods are hierarchically arranged in terms of assumptions and precision of inferences arising from the results. That is, assumptions and precision are relaxed as approaches move from equating to moderation. Mislevy's top level, equating, is the typical approach developers and researchers use to produce comparable forms. This level supports the finest distinctions in ability gradations. The methods evaluate test comparability through the use of statistical procedures in which comparisons are made directly between performances across forms. In addition to building forms from the same blueprints, the goal of content equivalence has typically been achieved by using identical development procedures, materials, and testing conditions. It is not clear whether this method of securing score consistency or equivalence is sufficient for producing forms with comparable inferences when forms include both standard and nonstandard versions. To date, it does not appear that other score equivalence methods have been considered to handle forms from the same blueprints for cases in which presentation or testing conditions are not identical.

UNCORRECTED PROOFS © AMERICAN PSYCHOLOGICAL ASSOCIATION

APA-HTA_V3-12-0603-018.indd 383 12/10/12 1:30 PM

Calibration, Mislevy's (1993) second level of linking, assumes that a well-specified content domain is the common frame of reference (e.g., content standards), and it evaluates the degree to which each form reflects that referent. The forms are compared with one another only indirectly. In development, calibration seems to assume that the forms do not use the same test specifications but substantively refer to the same referent throughout construction. As such, part of demonstrating adequate calibration will revolve around a quantified criterion estimate of the referent or detailed judgments from expert raters about the degree of alignment of the items on forms with the corresponding aspects of the target reference domain. Depending on the precision of analysis, comparisons may be made at the level of achievement standards, and possibly at some designations within the standards as well. Social moderation is the third level of linking in which the referent is levels of performance (e.g., the academic achievement levels). Here, forms are not designed to be parallel, and a looser form of expert judgment than calibration is utilized to evaluate how well the combined cognitive demand or other aspects of the content domain on each form supports comparability of performances. Empirical evaluations of linking in this case could compare the judgments about the forms, the subscore or total score performance of students, and perhaps some other independent judgments about the target abilities of the students. This level produces the least specific degree of comparability.

In 2006, federal funding was provided to continue to wrestle with comparability issues in statewide K–12 content testing when tests are given under varying conditions (Bazemore, 2006). This project used three general guiding questions to focus its work:

1. What do we want when we want score comparability?
2. What do we mean when we say comparability for a given purpose?
3. How can we evaluate comparability?

As the project unfolded, the questions were interpreted as follows (Winter, 2010a): The first question focused on the inferential achievement claims the test evidence can support. Documentation of the design of test development and subsequent procedures used to produce the evidence will need to pass scrutiny and should be evaluated through the lenses of appropriateness for capturing the knowledge and skills of particular students in particular situations. In other words, the evidence is viable if the logic of the overall design and individual procedural expectations can be argued through precedence to address and minimize alternative explanations, the implementation of the design and expectations are consistent with what is intended, and the implementation of the procedures themselves are implemented systematically and in a defensible manner. It is probable that test score evidence will come, to a reasonably large degree, from viable evidence at the item level, including systematic protocols and procedures associated with how some items responses are scored.

The second question addresses the level of comparability that is desired. For instance, is comparability focused at the achievement standards level (a series of about four school performance levels required under ESEA legislation for public school accountability), individual scale score level, or single cut-point level? This level of comparability makes a difference for the kinds of evidence that need to be collected, with the overall expectation that scores from both the general test and variation should be considered interchangeable enough and without flags. If the focus is one cut-point score (as in pass–fail), the whole assessment exercise should be focused on producing performances correctly identified on one side of the cutoff or the other. If more than one but a discrete number of scores are of interest, then interchangeability documentation needs to address the same question at each of the relevant scores. When raw or scale scores are the focus, then evidence needs to demonstrate that multiple scores along a continuous range are measuring similar enough knowledge or skills for the students taking each form. The third question focuses on how to analyze the evidence and make decisions about whether the evidence is good enough. Winter (2010a) has argued that there must be sufficient evidence of both content and construct equivalence and score equivalence, and that sufficient evidence along these lines form the basis of how one might judge the comparability of given materials for a given purpose.

UNCORRECTED PROOFS © AMERICAN PSYCHOLOGICAL ASSOCIATION

APA-HTA_V3-12-0603-018.indd 384 12/10/12 1:30 PM

Content and construct equivalence. The definition of content and construct equivalence as Winter (2010a) has applied this term focuses on grounding the score inferences across all variations considered to be interchangeable, in documented judgments and empirical evidence of the intended constructs being measured. Content and construct equivalence also involves ensuring that the user can have confidence that the meanings are the same (or the same enough). This aspect of equivalence reflects the analysis of evidence produced to defend the first question. Kopriva (2008) has argued that for equating, both adequate judgments and sufficiently rigorous empirical validation of the content and construct target equivalence need to undergird claims of score equivalence. Some elements of empirical support should supplement the judgments of content and construct equivalence at the other linking levels as well.

To make judgments about content and construct equivalence for ELs and SwDs, development methods designed to promote correspondence across items are referred to in earlier sections of this chapter. Back-translation and simultaneous (across languages) test development methods are important for ELs when the focus is content and construct equivalence between English and translated forms (e.g., see Ercikan, Gierl, McCreith, Puhan, & Koh, 2004; see also Chapter 26, this volume). Alignment analysis and other types of independent expert evaluations are examples of judgments that are also needed. For instance, judgment review procedures of item variations targeted to the same test specifications include those used by Gierl and Khaliq (2001), and alignment reviews such as those utilized by Webb, Alt, Ely, and Vesperman (2005) could be used to evaluate forms. Some researchers have used judgment techniques to evaluate the content similarity and comparability of cognitive complexity levels in items across forms (e.g., Kopriva, Wiley, & Winter, 2007; O'Neil, Sireci, & Huff, 2003–2004). Williamson, Bejar, and Sax (2004) explored how and when comparability might be affected when open-ended responses were scored using human and automated graders. After analyzing the judges' criteria for assigning scores and how the judges appeared to draw conclusions, they discussed how internal discrepancies might be handled to mitigate differences that arise.

For forms not built to be parallel, content experts may review the bodies of knowledge and skills assessed across forms and determine whether the same level of content complexity exists in both. Quality of judgments can be evaluated using statistics such as the confidence-interval approach proposed by Penfield and Miller (2004) or those used in standard setting. Approaches defined in multidimensional scaling or other similar content validation methods may also be appropriate to use in some situations (e.g., Haertel, 2003; Sireci, 1998).

Score equivalence. Score equivalence focuses on documenting that the scores from the variation and the general forms are behaving in the same way (or the same enough) for students with similar abilities. Evidence that will be analyzed for this aspect of equivalence comes from data that are appropriate to address the second question—that is, to defend the claims of interchangeability at the level of purpose. Examples of construct equivalent evidence that need to be evaluated include same-standards coverage, similar criteria for inclusion, similar judgments about relevant cognitive demands, and similar internal structure. Given evidence that data are drawn from samples for which similarity of student groups on important variables can be documented (e.g., through random assignment, control for differentiated ability using recognized methods, or evidence of similar distributions on relevant background variables), score equivalent evidence includes similar enough proficiency percentages, similar enough score distributions, similar enough structure of forms, and similar enough rank order. How "enough" is defined is a key part of determining score equivalence for particular uses and purposes.

When standard and nonstandard forms are designed to be parallel, statistical equating is the preferred approach to obtaining score equivalence because of the precision with which the equated scores can differentiate performance. Explanations of equating methods are outside the scope of this book. Basically, texts such as Kolen and Brennan (1995) have summarized a number of methods that collect test data from either equivalent or nonequivalent groups. When the distributions of groups are

UNCORRECTED PROOFS © AMERICAN PSYCHOLOGICAL ASSOCIATION

APA-HTA_V3-12-0603-018.indd 385 12/10/12 1:30 PM

considered to be equivalent (i.e., through random selection), linear equating and equipercentile techniques have been derived and similar techniques have been developed to handle nonequivalent groups as well. For most of these methods, data are collected on different forms or tests for the different groups. Most companies have moved to using item response theory techniques with nonequivalent groups to produce equated scores. This approach specifies that a subset of common items are given to the different groups as well as items that vary across groups. Item parameters for the common items are set across groups and maximum likelihood techniques are used to estimate the parameters for the rest of the items.

A number of different types of calibration and social moderation procedures have been identified in the past few years. Most often these look like modified standard-setting procedures, such as the Modified-Angoff and Bookmark methods (e.g., see Brennan, 2006; Cizek & Burg, 2005). See Feuer, Holland, Green, Bertenthal, and Hemphill (1999); Kopriva (2008); and Mislevy (1993), among others, for a more detailed discussion of this topic.

The Bazemore project yielded reports on a number of different studies that addressed methodological aspects of both content and construct equivalence and score equivalence (Winter, 2010b). These included findings about propensity score matching as an option to repeated measures methodology (Lottridge, Nicewander, & Mitzel, 2010); video versus paper-and-pencil forms for ELs of different languages, and factor analyses and multidimensional scaling to evaluate form differences (Sireci & Wells, 2010); simultaneous item codevelopment and anchor item methods (DePascale, 2010a); an evaluation model for a modified achievement test method for selected SwDs (DePascale, 2010b); and qualitative comparisons and judgment-based methods for students with significant cognitive disabilities and some ELs (Barton & Winter, 2010). Literature reviews reported in Winter (2010b) also summarized papers about paper-based versus computer-based modes of administration, translations and English forms for ELs, and plain-language editing.

CONCLUSION

Within the past 20 years, fields of cognitive psychology, educational practice, and accountability policy have each emphasized the diversity of the U.S. student population. One implication of this insight is that variable content testing methods predicated on making the same inferences about the content abilities of students tested under conditions that are designed to minimize challenges that are irrelevant to the academic knowledge and skills under scrutiny are most likely here to stay. To date, however, the assessment specialists, including academics and researchers, practitioners in state and local educational agencies, and test publishers, who design, implement, and interpret results from these testing systems, are far from unanimous as to how to construct, use, and defend content testing systems that include these variations. This chapter summarized how several of the relevant aspects of this important topic have been conceptualized to date and reviewed empirical work that has been completed to investigate the issues. Together this literature forms a body of work that defines the complexity of the topic and points to considerations for the future. The strength of the work is that it points to a multidimensional framework associated with effective academic measurement when systematic construct-irrelevant needs of certain populations would otherwise confuse the inferences about achievement that can be defensibly drawn for these groups. Although several next steps have been identified throughout the chapter, central to most of these is a focus on increasing the number of well-designed and thoughtfully implemented empirical studies to confirm or dispute hypothesized solutions. The finding from these investigations would, in turn, provide additional nuance, rigor, and direction to the discussion considered here.

References

Abedi, J. (2007, April). *Research on the validity of accommodations to help states with providing accommodations for ELL students.* Paper presented at the American Educational Research Association Annual Meeting, Chicago, IL.

Abedi, J. (2011). Assessing English language learners: Critical issues. In M. del Rosario Basterra, E.

UNCORRECTED PROOFS © AMERICAN PSYCHOLOGICAL ASSOCIATION

APA-HTA_V3-12-0603-018.indd 386 12/10/12 1:30 PM

Trumbull, & G. Solano-Flores (Eds.), *Cultural validity in assessment* (pp. 49–71). New York, NY: Routledge.

Albers, C. A. (2011). *The alternate ACCESS for ELLs with significant cognitive disabilities*. Madison: University of Wisconsin, Wisconsin Center for Education Research.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (3rd ed.). Washington, DC: American Educational Research Association.

Barton, K. E., & Winter, P. C. (2010). Evaluating the comparability of scores from an alternative format. In P. C. Winter (Ed.), *Evaluating the comparability of scores from educational achievement test variations* (pp. 95–104). Washington, DC: Council of Chief State School Officers.

Bazemore, M. (2006). *Funded proposal to the U.S. Department of Education, Office of Elementary and Secondary Education. Strengthening the comparability and technical quality of test variations*. Raleigh, NC: Department of Education.

Brennan, R. L. (Ed.). (2006). *Educational measurement* (4th ed.). Westport, CT: Praeger.

Callahan, R. M., & Gándara, P. (2004). On nobody's agenda: Improving English-language learners' access to higher education. In S. Michael (Ed.), *Teaching immigrant and second-language students: Strategies for success* (pp. 107–127). Cambridge, MA: Harvard Education Press.

Capps, R., Fix, M., Murray, J., Ost, J., Passel, J., & Herwantoro, S. (2005). *The new demography of America's schools: Immigration and the No Child Left Behind Act*. Washington, DC: Urban Institute.

Carr, T. G. (2008). Report of state specialists' review of STELLA decision-making algorithms. In T. Siskind (Ed.), *Final report for the AVAD grant* (pp. 2–6). Columbia: South Carolina Department of Education.

Carr, T. G. (2009, April). *It's about time: Matching English learners and the ways they take tests by using an online tool to properly address individual needs*. Paper presented at the National Council on Measurement in Education, San Diego, CA.

Christensen, L., Lazarus, S., Crone, M., & Thurlow, M. (2008). *2007 state policies on assessment participation and accommodations for students with disabilities* (Synthesis Report 69). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Christensen, L. L., Thurlow, M. L., & Wang, T. (2009). *Improving accommodations outcomes: Monitoring instructional and assessment accommodations for students with disabilities*. Washington, DC: Council of Chief State School Officers.

Cizek, G., & Burg, S. (2005). *Addressing test anxiety in a high-stakes environment strategies for classrooms and schools*. Thousand Oaks, CA: Sage.

Data Accountability Center. (2007). *Table 1–11: Children and students served under IDEA, part B, in the U.S. and outlying areas, by age group, year, and disability category-Fall 1998 through fall 2007*. Retrieved from https://www.ideadata.org/arc_toc9.asp#partbCC

DePascale, C. (2009). *Formative reform: Purposeful planning for the next generation of assessment and accountability systems*. Retrieved from http://www.nciea.org/publications/RILS_FormativeReform_CD09.pdf

DePascale, C. (2010a). Evaluating linguistic modifications: An examination of the comparability of a plan English mathematics assessment. In P. C. Winter (Ed.), *Evaluating the comparability of scores from educational achievement test variations* (pp. 69–94). Washington, DC: Council of Chief State School Officers.

DePascale, C. (2010b). Modified tests for modified achievement standards: Examining the comparability of scores to the general test. In P. C. Winter (Ed.), *Evaluating the comparability of scores from educational achievement test variations* (pp. 105–118). Washington, DC: Council of Chief State School Officers.

Elliott, S. N., Kratochwill, T. R., & Gilbertson-Schulte, A. (1999). *Assessment accommodations guide*. Monterey, CA: CTB McGraw-Hill.

Emick, J., & Kopriva, R. J. (2006, April). *Access-enhanced item development: A summary*. Paper presented at the American Educational Research Association, San Francisco, CA.

Emick, J., Wiley, D. E., & Kopriva, R. J. (2007, April). *The validity of large-scale assessment scores for ELLs under optimal testing conditions: Does validity vary by language proficiency?* Paper presented at the American Educational Research Association Annual Meeting, Chicago, IL.

Ercikan, K., Gierl, M. J., McCreith, T., Puhan, G., & Koh, K. (2004). Comparability of bilingual versions of assessments: Sources of incomparability of English and French versions of Canada's national achievement tests. *Applied Measurement in Education, 17*, 301–321. doi:10.1207/s15324818ame1703_4

Feuer, M. J., Holland, P. W., Green, B. F., Bertenthal, M. W., & Hemphill, F. C. (1999). *Uncommon measures: Equivalence and linkage among educational tests*. Washington, DC: National Academy Press.

Fields, R. (2008). *Inclusion of special populations in the national assessment: A review of relevant laws and*

UNCORRECTED PROOFS © AMERICAN PSYCHOLOGICAL ASSOCIATION

APA-HTA_V3-12-0603-018.indd 387 12/10/12 1:30 PM

regulations. Washington, DC: Report to National Assessment Governing Board.

Fry, R. (2007). *How far behind in math and reading are English language learners?* Washington, DC: Pew Hispanic Center.

Fuchs, L. S., Fuchs, D., Eaton, S. B., & Hamlett, C. B. (2005). *Dynamic assessment of test accommodations.* San Antonio, TX: Psychological Corporation.

Fuchs, L. S., Fuchs, D., Eaton, S. B., Hamlett, C. B., Binkley, E., & Crouch, R. (2000). Using objective data sources to enhance teacher judgments about test accommodations. *Exceptional Children*, *67*, 67–81.

Fuchs, L. S., Fuchs, D., Eaton, S. B., Hamlett, C. B., & Karns, K. M. (2000). Supplementing teacher judgments of mathematics test accommodations with objective data sources. *School Psychology Review*, *29*, 65–85.

Gándara, P., & Rumberger, R. W. (2009). Immigration, language, and education: How does language policy structure opportunity? *Teachers College Record*, *111*, 750–782.

García, E. E., Jensen, B. T., & Scribner, K. P. (2009). The demographic imperative. *Educational Leadership, 66*(7), 8–13.

Gierl, M. J., & Khaliq, S. N. (2001). Identifying sources of differential item and bundle functioning on translated achievement tests. *Journal of Educational Measurement, 38*, 164–187. doi:10.1111/j.1745-3984.2001.tb01121.x

Grant, R., & Cook, H. G. (2010). *Relating English language proficiency to mathematics performance through structural equation modeling.* Unpublished manuscript.

Haertel, E. H. (2003, April). *Evidentiary argument and comparability of scores from standard versus nonstandard test administrations.* Paper presented at the National Council of Measurement in Education, Chicago, IL.

Haertel, E. H., & Wiley, D. E. (2003, August). *Comparability issues when scores are produced under varying testing conditions.* Paper presented at the psychometric Conference on Validity and Accommodations, University of Maryland, College Park, MD.

Helwig, R., & Tindal, G. (2003). An experimental analysis of accommodation decisions on large-scale mathematics tests. *Exceptional Children*, *69*, 211–225.

Hernandez, D. J., Denton, N. A., & Macartney, S. E. (2008). Children in immigrant families: Looking to America's future. *Social Policy Report, 22*(3), 3–22.

Hipolito-Delgado, C., & Kopriva, R. J. (2006, April). *Assessing the Selection Taxonomy for English Language Learner Accommodations (STELLA).* Paper presented at the American Educational Research Association, San Francisco, CA.

Hollenbeck, K., Tindal, G., & Almond, P. (1998). Teachers' knowledge of accommodations as a validity issue in high-stakes testing. *Journal of Special Education*, *32*, 175–183. doi:10.1177/002246699803200304

Individuals With Disabilities Education Improvement Act of 2004, Pub. L. 108–446, 118 Stat. 2647 (codified at 20 U.S. C. § 1400 *et seq.*).

Kieffer, M. J., Lesaux, N. R., Rivera, M., & Francis, D. J. (2009). Accommodations for English language learners taking large-scale assessments: A meta-analysis on effectiveness and validity. *Review of Educational Research*, *79*, 1168–1201. doi:10.3102/0034654309332490

Kettler, R. J., Elliott, S. N., & Beddow, P. A. (2009). Modified achievement test items: A theory-guided and data-based approach for better measurement of what students with disabilities know. *Peabody Journal of Education, 84*, 529–551.

Kolen, M. K., & Brennan, R. L. (1995). *Test equating: Methods and practices.* New York, NY: Routledge.

Kopriva, R. J. (2008). *Improving testing for English language learners: A comprehensive approach to designing, building, implementing, and interpreting better academic assessments.* New York, NY: Routledge.

Kopriva, R. J., Cameron, C., & Gabel, D. (2010). *ONPAR findings from the 2008 science study for 4th and 8th graders.* Report from the University of Wisconsin Research Series. Retrieved from http://www.ONPAR.US

Kopriva, R. J., & Carr, T. G. (2009, June). *Building comparable computer-based science items for English learners: Results and insights from the ONPAR Project.* Paper presented at the National Conference on Student Assessment Annual Meeting, Los Angeles, CA.

Kopriva, R. J., Emick, J., Hildago-Delgado, C. P., & Cameron, C. (2007). Do proper accommodation assignments make a difference? Examining the impact of improved decision-making on scores for ELLs. *Educational Measurement: Issues and Practice*, *26*(3), 11–20. doi:10.1111/j.1745-3992.2007.00097.x

Kopriva, R. J., Gabel, D., & Cameron, C. (2011, April). *Designing dynamic and interactive assessments for English learners which directly measure targeted science constructs.* Presentation at the American Education Research Association Annual Meeting, New Orleans, LA.

Kopriva, R. J., & Hedgspeth, C. (2005). *Technical manual, Selection Taxonomy for English Language Learner Accommodation (STELLA) decision-making systems.* Madison: University of Wisconsin, Center for the Study of Assessment Validity and Evaluation.

UNCORRECTED PROOFS © AMERICAN PSYCHOLOGICAL ASSOCIATION

APA-HTA_V3-12-0603-018.indd 388 12/10/12 1:30 PM

Kopriva, R. J., & Lara, J. (2009). *Looking back and looking forward: Inclusion of all students in NAEP, U.S.'s National Assessment of Educational Progress* [commissioned paper]. Washington, DC: National Assessment Governing Board.

Kopriva, R. J., Lundberg, T., & Boals, T. (2011). *Performance, policy and politics: The place for well-reasoned accommodations in large-scale testing and accountability practices.* Manuscript submitted for publication.

Kopriva, R. J., & Mislevy, R. (2005). *Final research report of the Valid Assessment of English Language Learners Project* (C-SAVE Rep. No. 259). Madison: University of Wisconsin, Center for the Study of Assessment Validity and Evaluation.

Kopriva, R. J., Wiley, D. E., & Winter, P. C. (2007). *Analyzing skill complexity using specially constructed test scores.* Unpublished manuscript.

Koran, J., Kopriva, R. J., Emick, J., Monroe, J. R., & Garavaglia, D. (2006, April). *Teacher and multisource computerized approaches for making individualized test accommodation decisions for English language learners.* Paper presented at the National Council of Measurement in Education, San Francisco, CA.

Lottridge, S. M., Nicewander, W. A., & Mitzel, H. C. (2010). Summary of the online comparability studies for one state's end-of-course program. In P. C. Winter (Ed.), *Evaluating the comparability of scores from educational achievement test variations* (pp. 13–32). Washington, DC: Council of Chief State School Officers.

Mislevy, R. J. (1993). *Linking educational assessments: Concepts, issues, methods, and prospects.* Princeton, NJ: Educational Testing Service.

Mislevy, R. J. (1996). Test theory reconceived. *Journal of Educational Measurement, 33*, 379–416. doi:10.1111/j.1745-3984.1996.tb00498.x

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*, 3–62. doi:10.1207/S15366359MEA0101_02

No Child Left Behind Act of 2001, Pub. L. No. 107–110, 115 Stat. 1425 (2002). Retrieved from http://www2.ed.gov/policy/elsec/leg/esea02/107–110.pdf

O'Neil, T., Sireci, S. G., & Huff, K. L. (2003–2004). Evaluating the consistency of test content across two successive administrations of a state-mandated assessment. *Educational Assessment, 9*, 129–151.

Penfield, R., & Miller, J. M. (2004). Improving content validation studies using an asymmetric confidence interval for the mean of expert ratings. *Applied Measurement in Education, 17*, 359–370. doi:10.1207/s15324818ame1704_2

Pennock-Roman, M., & Rivera, C. (2006, April). *A review of test accommodations for ELLs: Effect sizes in reducing the mean achievement gap.* Paper presented at the American Educational Research Association, San Francisco, CA.

Perie, M., Grigg, W., & Dion, G. (2005). *The nation's report card: Mathematics 2005* (NCES 2006-453). Washington, DC: U.S. Government Printing Office.

Poteet, J. (1990). The what and how of modified assessment techniques. *Diagnostique*, *16*, 58–60.

Rigney, S., Wiley, D. E., & Kopriva, R. J. (2008). The past as preparation: Measurement, public policy, and implications for access. In R. J. Kopriva (Ed.), *Improving testing for English language learners: A comprehensive approach to designing, building, implementing, and interpreting better academic assessments* (pp. 37–64). New York, NY: Routledge.

Rivera, C., & Collum, E. (2006). *A national review of state assessment policy and practice for English language learners.* Hillsdale, NJ: Erlbaum.

Roach, A. T., & Elliott, S. N. (2006). The influence of access to the general education curriculum on the alternate assessment performance of students with significant cognitive disabilities. *Educational Evaluation and Policy Analysis*, *28*, 181–194.

Rumberger, R. W. (2006). *The growth of the linguistic minority population in the U.S. and California, 1980–2005*. Santa Barbara: University of California Linguistic Minority Research Institute.

Russell, M. (2010, October). *APIP.* Paper presented at the Reidy Interactive Lecture Series, Cambridge, MA.

Silver, D., Saunders, M., & Zarate, E. (2008). *What factors predict high school graduation in the Los Angeles Unified School District?* Santa Barbara: University of California.

Sireci, S. G. (1998). Gathering and analyzing content validity data. *Educational Assessment, 5*, 299–321. doi:10.1207/s15326977ea0504_2

Sireci, S. G., Li, S., & Scarpati, S. (2003). *The effects of test accommodations on test performance: A review of the literature* (Center for Educational Assessment Research Report No. 485). Amherst: University of Massachusetts, School of Education.

Sireci, S. G., & Wells, C. (2010). Evaluating the comparability of English and Spanish video accommodations for English language learners. In P. C. Winter (Ed.), *Evaluating the comparability of scores from educational achievement test variations* (pp. 33–68). Washington, DC: Council of Chief State School Officers.

Solano-Flores, G. (2010). Assessing the cultural validity of assessment practices: An introduction. In M. del Rosario Basterra, E. Trumbull, & G. Solano-Flores (Eds.), *Cultural validity in assessment* (pp. 3–21). New York, NY: Routledge.

UNCORRECTED PROOFS © AMERICAN PSYCHOLOGICAL ASSOCIATION

APA-HTA_V3-12-0603-018.indd 389 12/10/12 1:30 PM

Solano-Flores, G., & Trumbull, E. (2008). In which language should English language learners be tested? In R. J. Kopriva (Ed.), *Improving testing for English language learners: A comprehensive approach to designing, building, implementing, and interpreting better academic assessments* (pp. 169–200). New York, NY: Routledge.

Solomon, C., Jerry, L., & Lutkus, A. (2001). *The nation's report card: State mathematics 2000* (NCES 2001-519). Washington, DC: National Assessment of Educational Progress.

South Carolina Department of Education. (2010). *AccSelPro*. Retrieved from http://www.accselpro.org

Thurlow, M., Lazarus, S., Thompson, S., & Robey, S. (2002). *State participation and accommodation policies for students with disabilities: 2001 update*. Minneapolis, MN: National Center on Educational Outcomes.

Thurlow, M., Moen, R., & Wiley, H. I. (2004). *Biennial performance reports: 2002–2003 state assessment data*. Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Thurlow, M., Thompson, S., & Lazarus, S. (2006). Considerations for the administration of tests to special needs students: Accommodations, modifications, and more. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 653–673). Mahwah, NJ: Erlbaum.

Tindal, G., & Fuchs, L. (2000). *A summary of research on test changes: An empirical basis for defining accommodations*. Lexington, KY: Mid-South Regional Resource Center.

U.S. Department of Education. (2005). *Alternate achievement standards for students with the most significant cognitive disabilities: Nonregulatory guidance*. Washington, DC: Author.

U.S. Department of Education. (n.d.). *Building the legacy: IDEA 2004*. Retrieved from http://idea.ed.gov

Webb, N. L., Alt, M., Ely, R., & Vesperman, B. (2005, September). *The WEB alignment tool: Development, refinement, and dissemination*. Paper presented at the Council of Chief State School Officers' State Collaborative on Assessment and Student Standards, Technical Issues in Large-Scale Assessment Collaborative.

Weston, T. J. (2003). *NAEP validity studies: The validity of oral accommodations in testing* (NCES 2003-06). Washington, DC: National Center for Education Statistics.

Williamson, D. M., Bejar, I. I., & Sax, A. (2004). Automated tools for subject matter expert evaluation of automated scoring. *Applied Measurement in Education*, *17*, 323–357. doi:10.1207/s15324818ame1704_1

Winter, P., & Rabinowitz, S. (2006, May). *Two conditions necessary for comparability of achievement levels from test variations*. Paper presented at the Title I Peer Review Technical Assistance Meeting, U.S. Department of Education, Washington DC.

Winter, P. C. (2010a). Comparability and test variations. In P. C. Winter (Ed.), *Evaluating the comparability of results from educational achievement test variations* (pp. 1–12). Washington, DC: Council of Chief State School Officers.

Winter, P. C. (Ed.). (2010b). *Evaluating the comparability of results from educational achievement test variations*. Washington, DC: Council of Chief State School Officers.

Winter, P. C., & Gong, B. (2009, August). *Setting the stage*. Paper presented at the Test Score Comparability and Validity: Preparing for the Future of Assessment Dissemination Meeting, Washington, DC.

UNCORRECTED PROOFS © AMERICAN PSYCHOLOGICAL ASSOCIATION

APA-HTA_V3-12-0603-018.indd 390 12/10/12 1:30 PM