CHAPTER 9

UNIVERSAL TEST DESIGN

Looking Ahead in This Chapter

The modern classroom assessment approach of universal test design is presented. Universal design of assessment is based on the principle that all classroom assessment should be equally valid and accessible to every student. Validity and reliability issues relevant to universal test design are explored. Examples and guidelines for applying universal design in a real-life classroom are presented.

Objectives

After studying this chapter, you should be able to

- Define *universal design* and *universal test design*
- Summarize the principles of *universal design* and *universal test design*
- Provide examples of how universal design principles are applied in real classrooms
- Evaluate the extent to which universal design principles have been applied in a given classroom assessment

Stories From the Classroom

Ms. Clark Believes Variety Is the Spice of Life

Ms. Clark always appreciated a little variety in her life. After being in the same school for 5 years, Ms. Clark was excited to make the change to a new location, new students, and something a little different. She loved her old school and the students she got to work with there, but the new school was a little closer to home and would provide the opportunity to work with friends she has had in the district. Ms. Clark had also worked with the assistant principal in the past, training new teachers on the district plan to integrate Social Studies and English Language Arts, so she knew she would love the administration, too.

The diversity of the student population was not a surprise. Ms. Clark lived close to the school and knew the community well. There were even a few students who she saw regularly at the park near her house. She knew the material she was using at her old school would work just as well for the students in her new school. After all, it was the same district curriculum. There wouldn't be much new prep at all because she already had a bunch of assessments she had made and those had been perfected over the years.

At the end of every school year, teachers are asked to fill out a form with notes on each individual student. Teachers who will be getting the student the following year can then look at these files and plan accordingly. Ms. Clark was a little surprised when she noticed three intermediate level English Language Learners along with three additional students with minor disabilities who required instructional modifications. She thought she would need to teach a little differently to make sure everyone in the class had the opportunity to learn the material.

Two weeks had passed and Ms. Clark was happy with how the teaching had been going. It was time for the chapter test to really see how well those instructional changes had helped students learn the material. This took her a little off guard when some of her students appeared to have a really hard time taking the test. Looking over the completed tests, Ms. Clark decided that maybe her tests needed a little adjustment as well. Ms. Clark realized that maybe her assessments didn't work well for every student in her class. She remembered something from college about ways to ensure that tests worked equally well for every student, regardless of their characteristics. Maybe she still had some of her textbooks (though she had sold most of them back to the bookstore). When she got home, she started searching . . .

(To Be Continued)

> *"Universal design is . . . an enduring design approach that assumes that the range of human ability is ordinary, not special; . . . the experience of imaginative designers around the world reveals the range of applications that delight the senses and lift the human spirit when universal design is integral to the overall concept."*
>
> Elaine Ostroff, *Universal Design Handbook*
> (Preiser & Ostroff, 2001)

The opening quotation in this chapter describes a general design process that is wide-open and creative and that results in products useful to the fullest variety of people. The concept of **universal design** began as an architectural and engineering philosophy, but it has lately been embraced by teachers and other assessment designers. The idea, popularized by architect Ron Mace, who developed the nation's first state building accessibility code in North Carolina in the *1970s*, was to produce buildings and other physical environments so that they were accessible to all, including those who use wheelchairs or have other physical disabilities (Bowe, 2000, 2005). The underlying assumption of universal design is that all aspects of our world can be planned from the beginning to allow access and use by everyone.

Let's start with a concrete definition of what we are talking about:

Universal design the design of products and environments to be useable in a meaningful and similar way by all people.

The approach has spread to include other populations besides those with physical disabilities, such as the elderly, those with cognitive disabilities, and those whose primary language is not English, and to other areas, such as computer and website use, and education (Mcguire, Scott, & Shaw, 2006; Thompson, Johnstone, & Thurlow, 2002; Thompson, Johnstone, Anderson, & Miller, 2005). Those standards, which initially described a physical environment, have been applied to processes that include some physical aspects, but include experiential aspects as well. This chapter explores how this very modern concept of universal design can be applied to the very modern world of classroom assessment.

THE CASE FOR UNIVERSAL TEST DESIGN

Many of today's classrooms work very differently from the ones your parents remember, especially when it comes to assessment. It wouldn't necessarily be

clear, though, how different today's teacher-made tests might be because if you saw a quiz sitting on a teacher's desk, it would look at first glance like a "normal" everyday classroom test. It's more and more likely, however, that that multiple-choice quiz has been designed from the start to be meaningful, valid, and reliable for *all* students. It began with the teacher's choice of what to measure. He or she probably carefully defined the assessment objectives to match important instructional objectives and did not assign points for irrelevant knowledge or skills. Then, the teacher likely paid careful attention to the test instructions. He or she worded the directions simply, not assuming that students would remember or completely understand the rules from previous tests. Many modern teachers also do something on their assessments that few teachers did in the past. They give an example of an item and show how to respond correctly. The items themselves are written in clear language, and unless there is a reason for doing it otherwise, the vocabulary level is no more advanced than necessary. These choices remove some of the language obstacles that can affect performance for some of their students. Another concern that many of today's teachers have is the actual printed formatting regarding font size, the use of lots of "white" space (the space on the test that is blank), the use of capitals, and so on. This makes things easier for students who may have vision or perceptual difficulties related to learning disabilities. These strategies, along with other simple choices many teachers make, help produce assessments that are consistent with the principles of universal design.

Universal design of assessment is part of the accessibility movement in the United States, and the common acceptance of the approach is the result of several political and legislative developments over the last half century. An outline of the brief history of universal testing highlights these developments:

- 1950s. Disabled veterans and advocates for people with disabilities demanded opportunities in education, employment, and housing. *Barrier-free Movement* began.
- 1960s. American Standards Association (ANSI) published building accessibility standards. By the end of the decade, most states had adopted ANSI standards. *Architectural Barriers Act of 1968* required that all federally funded buildings be made accessible.
- 1970s. Architect Ron Mace popularized *universal design* concept. *Section 504*, *Rehabilitation Act of 1973*, outlaws discrimination on the basis of disability. *Education for Handicapped Children Act of 1975* (IDEA) guarantees a free and appropriate education for all children with disabilities.

- 1980s. *Fair Housing Amendments Act of 1988* required that most housing be accessible to those with disabilities.
- 1990s. *Americans with Disabilities Act* (ADA) banned discrimination in employment and required full access to, and use of, virtually all places and services. Center for Universal Design suggested seven *Universal Design Principles*.
- 2000s Educators began to develop classroom assessment methods that apply the principles to assessment.

Universal test design, or *universal design of assessment*, is the most modern of the approaches presented in this book. As a coherent philosophy and set of guidelines, universal assessment is less than 20 years old and is even now developing. Both classroom teachers and standardized test developers have begun to explore design of assessments that work equally well for all students regardless of their characteristics. Those who use the jargon of measurement would say, more specifically, that assessments should work equally well for all students regardless of their *construct-irrelevant characteristics*. They should be valid for all.

Principles of Universal Design

There are seven broad, established standards for universal design (Center for Universal Design, 1997):

1. Equitable Use
2. Flexibility in Use
3. Simple and Intuitive Use
4. Perceptible Information
5. Tolerance for Error
6. Low Physical Effort
7. Size and Space for Approach and Use

Each of these general standards has been interpreted in a more specific assessment context. The goal of universal design in assessment would be to "allow participation of the widest range of students, and to (produce) valid inferences about performance for all students who participate in the assessment"

Universal design advocates believe that your teaching and your assessment should be open to everyone.

(Thompson et al., 2002, p. 7). Though no assessment will be completely accessible or valid for all, the objective is to be as inclusive as possible. Some of these interpretations match directly to the broad seven standards, and some do not:

1. Inclusive assessment population
2. Precisely defined constructs
3. Accessible, nonbiased items

4. Amenable to accommodations
5. Simple, clear, and intuitive instructions and procedures
6. Maximum readability and comprehensibility
7. Maximum legibility

So what do these standards look like when they are applied to classroom assessments? A teacher-developed classroom assessment built under the philosophy of universal design may on its face look somewhat similar to an assessment from 20 years ago, though there will likely be some technical differences (e.g., larger fonts, more white space) that could be noticed, and the wording of directions and questions may be simplified. The bigger differences, however, are likely to be in the choice of tasks, questions, and administrative procedures, as well as in the planning.

Table 9.1 presents the broad universal design standards and the much more specific universal test design standards and indicates how the latter standards are faithful to the former standards. It also describes in detail what the observable characteristics of a classroom assessment are that have been developed under this philosophy. Because there is an emphasis on precisely defined assessment targets, it is a goal of universal design that performance should not be based on anything other than those assessment targets (AERA, APA, & NCME, 1999) and that points are awarded for knowledge or performance, not other abilities (e.g., speed, handwriting, perhaps spelling and grammar). There also is a focus on eliminating cultural bias. In this context, items are considered biased if groups of equal ability have different probabilities of answering them correctly. A third goal for development of these universally accessible assessments is that if accommodations are necessary (if a different form or format of the assessment must be used because of a student's disability), adapting the assessment into a more accessible format is as easily accomplished as possible. Although it is the ultimate principle of universal test design that the same test can be used by all (in the same way that the same entrance to a courthouse can be used by all), accommodations will occasionally still be necessary. Universal design plans for that up front. This means, for example, that graphical material should be easily describable to a blind student or, perhaps, graphical material should not be used at all unless it is vital to the question or assessment task. The remaining applications of universal design for assessment describe various physical and formatting aspects of a test.

Table 9.1 Applying Universal Test Design Principles

Universal Design Principles								
Equitable Use	Flexibility in Use	Simple and Intuitive Use	Perceptible Information	Tolerance for Error	Low Physical Effort	Size and Space for Approach and Use	**Universal Test Design Principles**	**Observable Characteristics**
X	X						1. **Inclusive assessment population** Opportunity for participation for all members of the target population regardless of physical characteristics, culture, linguistic background, or cognitive abilities.	This information is difficult to "observe," but most teacher-developed assessments are consistent with this principle.
X				X			2. **Precisely defined constructs** Performance should not be affected by construct-irrelevant variance, processes that are extraneous to the intended construct (AERA et al., 1999).	Points awarded for knowledge or performance, not construct-irrelevant tasks (e.g., speed, handwriting, perhaps spelling and grammar). Wording for math problems should be simple and clear.
X	X						3. **Accessible, nonbiased items** Items are biased if groups of equal ability have different probabilities of answering correctly. Items also should be free of culturally offensive content.	Words, phrases, and concepts are commonly used across cultures and languages. No pop culture references (e.g., TV, music). No stereotypes or offensive terms.

Universal Design Principles								
Equitable Use	Flexibility in Use	Simple and Intuitive Use	Perceptible Information	Tolerance for Error	Low Physical Effort	Size and Space for Approach and Use	**Universal Test Design Principles**	**Observable Characteristics**
X	X		X			X	4. **Amenable to accommodations** The way in which a test is presented can easily be changed to remove unintended disadvantages for English language learners or for those with disabilities.	Horizontal text. No construct-irrelevant graphs or pictures. Graphics are simple and clear. Keys and legends at top or right of item. No time limits. Subsections of tests are independent of each other.
X		X		X			5. **Simple, clear, and intuitive instructions and procedures** "Assessment instructions and procedures need to be easy to understand, regardless of a student's experience, knowledge (or) language skills . . ." (Thompson et al., 2002, p. 14)	Consistent instructions (e.g., circling correct answer). Directions allow students to work independently without questions. Practice or sample items are provided. Numbered items.

A Closer Look

Great Minds Think Differently

Universal test design is partly derived from a broader educational approach, *Universal Design for Learning*, which was developed to embrace the differences in students, including the different ways that people think, organize their thoughts, and learn new information and skills (Dolan & Hall, 2001). Because of modern brain imaging technology, we can now almost literally see thinking take place, and we know quite a bit about how different parts of the brain oversee different activities. It turns out that there are large differences from person to person in the nature of these different brain areas, how directions are understood, and how the brain is networked to complete simple and complex tasks. Each student brings a unique mix of strengths, challenges, and preferences to the learning environment, and it is less common these days for teachers to think of there being a few types of students or learning styles. Essentially, brain regions can be grouped into three types of networks that control learning:

- **Recognition**. This network specializes in receiving and analyzing information. It processes the content, identifies what is new, what is already known, and what is similar to current knowledge and skills, and it begins to organize the information.
- **Expression.** This type of processing plans and executes actions. If learning is the result of behaviors, it is this network that regulates those behaviors. Similarly, when taking a test or engaging in assessment tasks, it is the expression network that makes the decisions on following instructions.
- **Engagement.** This network is the affective, motivational, and attitudinal specialist. This system sets priorities and controls the energies allotted to engaging in learning or trying one's best on an assessment.

The tools, strategies, and technologies that support universal design of assessment are meant to flexibly meet the needs of all students, regardless of the ins and outs of how they individually happen to process information, engage in learning, and behave when assessed.

Technology

Universal Design Technology

New computer technologies that support universal test design allow for multiple modes of representation (Rose, Hall, & Murray, 2008; Salend, 2009) for taking in information and directions and responding with a test answer or assessment performance. For example, computers can give students control over font, size, and

color of a traditional test (which we can no longer correctly label as paper-and-pencil). Instructions, tasks, items, and assignments can be presented orally or in braille. Technology exists to turn text into speech and speech into text, and text and pictures can become touchable (Dolan & Hall, 2001).

The National Center on Universal Design for Learning manages a website that gives detailed examples of technological supports for universal design of classroom assessments. Many of these tools are computer-based, some are traditional hardware solutions, and some are guidelines for teachers designing their own tests.

Designers of computer-based assessments have given much thought to the application of universal design principles.

Technology for increasing the number of ways that information can be represented is described here:

http://www.udlcenter.org/aboutudl/udlguidelines/principle1

Tools for allowing for multiple means of expressing what students have learned are here:

http://www.udlcenter.org/aboutudl/udlguidelines/principle2

Ways of engaging all students in assessment regardless of their individual characteristics are discussed here:

http://www.udlcenter.org/aboutudl/udlguidelines/principle3

WHAT WE KNOW ABOUT UNIVERSAL TEST DESIGN

There is very little real research on the effect of universal design for teacher-made classroom assessments. Recommendations for the approach are driven primarily by theory and philosophy. That doesn't mean that universal design won't increase the usefulness of classroom assessments and allow for fair access to tests by more populations; it only means that it hasn't been studied much and we do not yet know for sure whether it makes a difference regarding assessment and learning. There are a few studies of universal design principles and their effect on *standardized* test performance, however. Much of the

research effort in universal design of assessment has focused on these formats because of the federal mandate to include all populations in statewide testing. Presumably, if application of universal design guidelines positively affects performance on standardized tests, the same applications should also increase performance on *classroom* assessments.

A second line of research related to universal design is the willingness of teachers to buy into the philosophy and apply its principles. Lombardi and Murray (2011) conducted a large survey of college faculty at one university about their attitudes toward the principles and instructional behaviors and expectations consistent with universal design. College teachers make many of the same decisions about their teaching as elementary and secondary teachers do, and their beliefs and understanding of this approach likely mirror those of K–12 instructors. The researchers found that teachers who were female, were newer on the job, or had been trained to teach students with disabilities felt much more positively toward minimizing barriers, adjusting assignments and requirements, providing easier access to course materials, and other universal design characteristics.

DOING A GOOD JOB OF ASSESSING ALL STUDENTS

A fairly modern validity concern with teacher-made or standardized tests is the validity of inferences made from such tests for students whose first language is not English. A second somewhat more traditional concern is the validity of these assessments for students with disabilities. Universal design is meant to respond to those validity concerns by producing assessments that not only fairly assess those students, but fairly assess all students regardless of their irrelevant characteristics. Beyond these concerns, though, there is a modern concept of validity that is frequently cited by supporters of the universal design approach. This aspect of validity is known as social consequences validity or *consequential validity*. The usefulness of an assessment is not only whether the test score accurately represents a particular domain of knowledge or skill, but includes whether the use of an assessment is *fair* and *just* in a social sense. The underlying argument for the need for universal design considerations is that traditional assessment scores may represent something a bit different for each student. If some of the variability in scoring is our old friend, *construct-irrelevant variance*, then the validity of those scores is questionable. If assessments are designed from the beginning so that all items are free from cultural bias, all students understand directions, all students can read and comprehend all items, and all students are capable of performing all assessment tasks, then construct- irrelevant variance is minimized.

Real-World Choices Teachers Make

Deciding the Purpose of a Test

Modern classroom teachers sometimes choose to view the validity of assessments as something a bit more than simply whether they measure what they are supposed to measure. They think of the social effect of their assessments on students. Samuel Messick (1993), a measurement philosopher, first suggested this idea of *consequential validity*. He pointed out that an assumption underlying the broad concept of validity is that tests should serve the purposes for which they are intended. If a teacher, school system, or state believes that the use of an assessment will ultimately help those involved by improving instruction, for example, or by increasing student learning, that intent becomes part of the validity requirement for the assessment. "Judging validity in terms of whether a test does the job it is employed to do—that is whether it serves its intended function or purpose—requires evaluation of the intended or unintended social consequences of test interpretation and use," he argued (Messick, 1993, p. 84). The educational and psychological measurement field incorporated Messick's arguments into a modern definition of validity which now "officially" stands as *the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests* (AERA et al., 1999, p. 9). Teachers with this view of validity are often concerned with the instructional time taken up by assessments, the effects of labeling on students, whether tests are biased, and other issues regarding the consequences on students from assessment.

A Changing Definition of Validity?

Good Question!

Q: So you just dedicated about eight chapters to the importance of validity defined as "whether the test measures what it is supposed to" and divided validity into three types: content, criterion, and construct. Now, you casually mention in a **Real-World Choices** box that validity means something else? What's up?

A: Yes, measurement folks now understand validity as defined by the phrasing in the *Standards* produced by the key relevant scientific professional organizations and quoted in this chapter. It is consistent with the modern perspective that validity refers to both the score of a test and the purpose of a test and that

validity is a single thing (without three different "types"). For classroom teachers, or anyone making decisions when developing or evaluating an assessment, however, it is still very useful to separate the various types of validity evidence and validity arguments into the three moderately distinct categories: content, criterion, and construct. Strategically focusing on each of the three categories of evidence as relevant to particular validity concerns is a useful way to clarify one's own convictions, philosophies, and understanding of teacher-designed tests.

Because universal test design is an approach, a concept, that can be applied to either traditional paper-and-pencil assessment or performance assessment, there are no particular special issues of reliability. One benefit of universal design, though, should to some extent have relevance to inter-rater reliability concerns. The level of subjectivity in any scoring system affects inter-rater reliability, and one source of subjectivity is bias. Evaluating unexpected responses or dealing with task performance that does not seem to meet assessment instructions or requirements is difficult. Responses will be more uniform when directions and tasks are described using text that is easily understood by all students. The range of performances should more closely match the rubric categories and expectations when the assessment is planned from the start following universal design guidelines. So one might expect less subjectivity in scoring when this modern approach is followed.

WHAT UNIVERSAL TEST DESIGN LOOKS LIKE IN THE CLASSROOM

> *"I'm not sure it's possible to create anything that's universally usable. It's not that there's a weakness in the term. We use that term because it's the most descriptive of what the goal is."*
>
> Ron Mace, Architect (1947–1998)

Let's begin figuring out how to apply the universal test design principles to classroom assessment by examining them in greater detail. Table 9.2 provides the goals principles with some useful "sub-principles" from the Center for Universal Design (1997) that help provide a transition from a theoretical approach to actual teacher choices and strategies. With this table we can begin

Table 9.2 Application Guidelines for Universal Test Design

Principles	Guidelines (Sub-principles)
1. **Equitable Use** The design is useful and marketable to people with diverse abilities.	Provide the same means of use for all users: identical whenever possible; equivalent when not. Avoid segregating or stigmatizing any users. Provisions for privacy, security, and safety should be equally available to all users. Make the design appealing to all users.
2. **Flexibility in Use** The design accommodates a wide range of individual preferences and abilities.	Provide choice in methods of use. Accommodate right- or left-handed access and use. Facilitate the user's accuracy and precision. Provide adaptability to the user's pace.
3. **Simple and Intuitive Use** Use of the design is easy to understand, regardless of the user's experience, knowledge, language skills, or current concentration level.	Eliminate unnecessary complexity. Be consistent with user expectations and intuition. Accommodate a wide range of literacy and language skills. Arrange information consistent with its importance. Provide effective prompting and feedback during and after task completion.
4. **Perceptible Information** The design communicates necessary information effectively to the user, regardless of ambient conditions or the user's sensory abilities.	Use different modes (pictorial, verbal, tactile) for redundant presentation of essential information. Provide adequate contrast between essential information and its surroundings. Maximize "legibility" of essential information. Differentiate elements in ways that can be described (i.e., make it easy to give instructions or directions). Provide compatibility with a variety of techniques or devices used by people with sensory limitations.
5. **Tolerance for Error** The design minimizes hazards and the adverse consequences of accidental or unintended actions.	Arrange elements to minimize hazards and errors: most used elements are the most accessible. Hazardous elements eliminated, isolated, or shielded. Provide warnings of hazards and common errors. Provide fail-safe features. Discourage unconscious action in tasks that require vigilance.
6. **Low Physical Effort** The design can be used efficiently and comfortably and with a minimum of fatigue.	Allow user to maintain a neutral body position. Use reasonable operating forces. Minimize repetitive actions. Minimize sustained physical effort.
7. **Size and Space for Approach and Use** Appropriate size and space is provided for approach, reach, manipulation, and use regardless of user's body size, posture, or mobility.	Provide a clear line of sight to important elements for any seated or standing user. Make reach to all components comfortable for any seated or standing user. Accommodate variations in hand and grip size. Provide adequate space for the use of assistive devices or personal assistance.

Source: Adapted from Center for Universal Design, 1997.

to picture what the somewhat abstract principles look like when applied to classroom assessment. What does *Flexibility in Use* mean in classroom assessment terms? Among other things, it means designing the layout of a pencil-and-paper test so that it is easily used by both right-handers and left-handers. What does *Tolerance for Error* look like? As the table tells us, one way it might manifest itself in classroom assessment is when test instructions include "warnings" to avoid common mistakes. The sub-principles are kind of like instructional behavioral objectives; they are more concrete ways to operationalize conceptual goals.

Layout and Format of Universally Designed Tests

While the sub-principles provide a good starting point, more is needed for a teacher who wishes to develop an assessment from scratch that meets the universal design guidelines. For an evidence-based list of assessment construction decisions that are effective and consistent with universal design principles, the summary of dozens of research studies compiled by Thompson et al. (2002) is indispensable. Highlights of their cataloging of research findings are presented here, but serious universal design test-makers should get their hands on a copy of the full "manual." (There is a free online copy at www2.lexcs.org/osep/pdf/Universal_Design_LSA.pdf and many other locations.) The specific technical suggestions and construction guidelines that follow are from their review and supported by empirical research.

Text Formatting. For Western-style readers, text that is flush to the left margin is easiest to read. "Fully justified" text (which is spaced so that both the left and right margins are flush or straight) is difficult for even expert readers to handle. Students recall text better when it is left justified and "ragged" (unjustified) on the right.

Type Size. Certain types (what we nonprinting experts call *fonts*) are universally better than others: 14-point type is better than 12 or 10 and has been shown to actually increase tests scores for all students. Students with moderate visual impairment require at least an 18-point size type.

Use at least 12-point for titles, footnotes, and such on graphics and tables. Fixed-space fonts are more legible than proportional-spaced fonts and those that are serif (have the tiny perpendicular lines at the end of each stroke, e.g., Times New Roman or `Courier`) may work better than sans serif fonts (types without those little lines, e.g., Arial or Helvetica).

Formatting Text. It is generally best to use standard typeface with the correct use of upper- and lowercase. This is generally more readable for all students than all uppercase or italicized texts. To emphasize some text, bolding works well. It works better than switching to all caps.

Text Line Length. The distance between the left and right margin of text makes a difference for some students. One study suggested that 4 inches is the "best" length for a line on a test form. Another suggested that about 12 words is a good maximum length. Whatever the length, if the lines are longer in inches, the font should be larger.

White Space. A large amount of space on a page that is blank without text or pictures can increase legibility and aid in directing focus to the text. (A better term for *white space* is *blank space* because paper and computer screen backgrounds are often not white.) A traditional rule of thumb for test construction is that half the page should be blank. Younger students appreciate even more white space. If this sounds like a lot of white space, examine the page in this book you are reading right now. Notice the blank space in the margins and between the lines of text and so on. The ratio of text and graphics compared with blank space is probably about 50/50 and it doesn't look odd.

Leading. The amount of space between lines matters. A lead space that is too short makes reading difficult for those with low vision or certain learning disabilities. Here are some rough guidelines:

- 12-point type, 2 to 4 points of leading
  - This is 12-point type with 2 points of leading.
- 14-point type, 3 to 6 points of leading
  - This is 14-point type with 3 points of leading.
- 16-point type, 4 to 6 points of leading
  - This is 16-point type with 4 points of leading.
- 18-point type, 5 to 6 points of leading
  - This is 18-point type with 5 points of leading.

Contrast. Some research has been done on legibility as it relates to the contrast between the text and the background. Use off-white or light pastel, nonglossy paper to prevent glare. Type should be black.

Bubble Answer Options. There are several points of consideration when the answer format for a traditional paper-and-pencil test requires that students fill in bubbles. Tests with small bubbles are difficult for students with low vision or trouble with fine motor skills. Students with learning disabilities may also benefit from larger text and answer marking options. There are mixed results in the literature as to negative consequences of having the answer sheet separate from the test (it may vary with student developmental level), but fewer errors are made when one can respond on the test itself.

Graphs and Tables. While black and white is generally the best color scheme for graphs and tables, symbols need to be particularly clear and distinct when color is not used. Use labels directly on graphs and maps, instead of using a separate legend or key elsewhere. Avoid using varying degrees of grey scale in tables and figures if perceiving the differences is meant to provide information.

Illustrations. For some students, illustrations cause a problem in visual discrimination or competition for attention between picture and text. Black-and-white drawings will be clearest to all. Illustrations may be in color to attract appropriate attention, but avoid green and red combinations because some students may have color blindness. Graphics should be right next to the relevant question or text. Some illustrations are merely decorative or entertaining (e.g., a Ziggy cartoon about how tough tests are), whereas others are integral to the assessment task or question. Consider using only those illustrations that are necessary to the assessment, because some students are better than others at knowing what to ignore and what is important.

Writing Universally Designed Tests

The science of universal test design has to do with the physical characteristics of a test that follow the key principles. The art of universal test design comes into play in the actual writing of an assessment. It is word choice in items, directions, and the terms used on an assessment that may lead to the dreaded *construct-irrelevant variance* in the scores for some students. Fortunately, researchers have suggested guidelines to follow when composing items and assessment tasks and when formulating directions.

General guidelines for knowing whether the content of an assessment follows universal design principles and allows "access" to all students are provided by Rakow and Gee (1987).

1. All students would likely have the experiences and prior knowledge necessary to understand the question.
2. The vocabulary, sentence complexity, and required reasoning ability are appropriate for all students' developmental levels.
3. Definitions and examples are clear and understandable.
4. Relationships are clear and precise.
5. Item content is well organized.
6. The questions are clearly framed.
7. The content of items is of interest to all students.

The wording used in assessments can make a difference. Brown (1999) suggests a variety of ways to ensure that assessments are written in "plain language." First, shorten the length of sentences wherever possible. Reduce needless wordiness and irrelevant text; break complex sentences into several shorter sentences. Second, unless it is important to use the jargon of a field, replace unusual words with more common synonyms. Brown gives as an example to say *use* instead of *utilize*. Avoid ambiguous words. Use proper nouns only when necessary. Third, be consistent across assessments and within each assessment. This means use the consistent graphic, table, and mapping conventions. Use the same font and layout every time. Use the same word for an important concept each time you use it. Finally, number or identify in some way each question. When asking related questions or questions with many parts, clearly mark each part with a bullet, letter, or number. Brown found that when students actually know the answers or have the assessed skill, they perform higher on plain language tests, but that performance was not affected for those who did not have the knowledge or skill. This is a good indication that the use of plain language tests affects only the construct-irrelevant variance in performance and increases validity. It increases fairness without disadvantaging any students.

Gaster and Clark (1995) have a list of guidelines for increasing readability, which are similar to Brown's. Among their somewhat more specific recommendations are these:

- When technical terms must be used, define them carefully and clearly.
- When breaking up compound sentences, state the most important idea first.
- Introduce one idea at a time, and develop more complex ideas logically.

- Make it clear to whom or what a pronoun refers.
- If time or setting are important, place them at the start of the sentence.
- Sequence steps in the exact order in which they should occur or be done.

Steps for Assessment Design

A detailed example of procedures for developing classroom assessments that follows the principles of universal design is provided by Ketterlin-Geller (2005). Though her example is specifically for designing a computer-enhanced assessment, the principles and applications generalize well to traditional paper-and-pencil tests. The assessment that is described is a 3rd grade math test. As the author points out, many of the procedures and development strategies used in this test are similar to those for other classroom assessment approaches. The difference is the "conscious and deliberate consideration of individual needs" along the way (p. 11).

Step 1. **Identify and define the construct.** What skill, ability, attitude, or knowledge domain is meant to be assessed? In Ketterlin-Geller's example, the construct was mathematical ability. More specifically, the construct was the knowledge and skills identified as standards for the 3rd grade in the state in which the assessment was developed. These were measurement concepts, geometry, probability, statistics, algebra concepts, calculation skill, and estimation skill.

Step 2. **Identify and define the population.** In this example, the population was all 3rd graders. This population included students with a wide variety of disabilities, linguistically diverse students and students, with a wide variety of cultural characteristics and cognitive abilities.

Step 3. **Choose the testing "platform."** Will it be traditional paper-and-pencil, performance assessment, computer-based, or some other assessment environment? At this step, the designers decided that they wanted flexibility in the level of support (e.g., practice items, navigation options, concentration aids, text-to-speech capability) and chose a computer environment.

Step 4. **Choose the item format.** Ketterlin-Geller and colleagues wished to use a traditional multiple-choice format. To increase reliability, they used five answer options instead of four (this reduces the likelihood of randomly guessing the correct answer). A left-to-right layout was chosen (question on left; answer options on right). Answer options were vertical, one beneath the other, which is consistent with universal design guidelines. Because the answer options would be indicated on a computer screen, they did not need to be labeled with A's, B's, C's, and so on. Because some students might have physical

disabilities, more than one way of indicating the correct answer was available (using a mouse or the keyboard). So that difficulties with attention and concentration were less likely to affect performance, the interface was designed so students could select an answer and review it as long as they wished before submitting it.

Step 5. **Compose and sequence the test.** This computerized, 3rd grade math test was written so that directions, prompts, and questions were simplified (the text is easy, not the difficulty level). In an example item provided (p. 15), a two-color graphic is shown of 11 circles. Each circle either is striped or has a crossed-lines pattern (crosshatch). Four of the circles are striped. The question is worded in a straightforward manner without superfluous text: "What is the probability of picking a striped ball?" Because the question is designed to assess understanding of probability concepts, and not geometry terms or anything else, the simpler word *ball* can be used instead of circle. The word *probability* should be used instead of a simpler word, though, because it is terminology central to the targeted skill. Answer options are succinct and only provide information necessary to answer the question: for example, "4 out of 11."

Step 6. **Finalize accommodation options.** This example had built-in accommodation options, such as text-to-speech options, which were available by clicking on a "speaker" icon. Students could listen to a question or directions as often as they wished. An alternative form was available with the same math questions in an even more simplified format. Access to the alternative form was automatic based on a brief pretest screening of sorts that assessed reading ability.

The author emphasizes that though this particular case example used computers for administration, the principles applied here can also be applied to traditional paper-and-pencil teacher-made classroom assessment. This is true, of course, as most universal design "rules" apply to wording of items, the layout of test components, and the up-front careful definition of the intended construct for assessment.

Examples of Universally Designed Directions and Items

The Kansas State Department of Education (2010) produces *Kansas Computerized Assessment* (KCA) tests for reading, math, and other subjects administered to Grades 3 through 11. They were developed following universal design guidelines. The publicly available test manuals include item examples, accommodation rules, and the exact directions meant to be read aloud to all

students. The tests are administered on computer and include online tools for crossing off answer options and erasing those marks. The directions and examples shown here are taken from the 2010–2011 Kansas Assessment Examiner's Manual (Kansas Department of Education, 2010). The sample items are taken from free, publicly available software that allows for student practice and can be found at http://www.cete.us/kap/downloads/downloads_kca_windows.htm.

Directions (to be read to students aloud)

"Try to answer all questions, even if you have to guess. If you are not sure about the correct answer . . . cross out any answers that you think are not correct. Choose the answer that you think is best. It is important to answer all questions. Does anyone need scratch paper?

"The questions in this test are multiple-choice. There is one correct or best answer to each question. Carefully read the question. Work the problem. You may use scratch paper . . . Decide which answer is correct or clearly better than the other choices.

"You are to complete the questions in each part as directed. When you have answered the last question, raise your hand, and I will verify that all of the questions have been answered. You may use [a] calculator on this part of the test."

Items

7th grade science item

1. Roberta wants to measure the mass of a ball of foil she has made from aluminum wrappers. It is about the size of an orange. What instrument would she use to measure its mass?
   - A spring scale
   - A graduated cylinder
   - A meter stick
   - A balance

11th grade social studies

4. Due to the Miranda decision, which right are police required to inform citizens they possess?
   - Right to trial by jury
   - Right to notify family
   - Right to an attorney
   - Right to confront witnesses

4th grade mathematics

1. A company made 3000 cards for a game. Ed has collected 2,419 cards so far. How many cards must Ed collect to have all 3,000?

   (Hint: You can use the calculator.)

   - 5,419 trading cards
   - 1,419 trading cards
   - 691 trading cards
   - 581 trading cards

Real-World Choices Teachers Make

Item Writing and Universal Test Design

The universal test design approach provides general guidelines for producing items that are fair for all students. This is much more of an art, of course, than a science. As an illustration, let's examine the items presented as examples in this chapter of questions, which are consistent with universal design principles. These are the sample items supplied by the State of Kansas for students who wish to practice for the state assessments. Kansas produced their item pool following universal design principles. Overall, these items are very consistent with the approach, of course. Wording is simple without superfluous text. There is plenty of "white space." Online tools and accommodations are built into the assessment. But, if one got picky, one could suggest ways that even these items might work better (at least using universal design as the criterion for "work better").

In the first item, the 7th grade science question, the term "foil" is used in one part of the question, while the term "aluminum" is used later. Universal design guidelines suggest using consistent terminology throughout a test (and, certainly, an item). In the next item, the 11th grade social studies question, the stem asks about which rights police are required to inform citizens they "possess." The simpler word "have" means the same thing and is at a lower level of vocabulary. (Recall the suggestion earlier that *utilize* means *use*, so one should just say *use*.) The final example provided in this chapter, the 4th grade math item, uses an applied math example involving a game involving cards (e.g., *Magic*, *Pokémon*, and other "trading card" games). It is likely that boys have more experience with this sort of real-world problem than girls do, so from a universal design perspective, girls have less "access" to this item.

Figure 9.3 Universal Test Design Assessment Instrument

Universal Design Elements	Score		
Accessible, nonbiased items Items are biased if groups of equal ability have different probabilities of answering correctly. Items also should be free of culturally offensive content. • Words, phrases, and concepts are commonly used across cultures and languages. • No pop culture references (e.g., TV, music, movies), idioms, colloquialisms. • No stereotypes or offensive terms.	**0** More than 1 violation	**1** 1 violation	**2** No violations
Amenable to accommodations The way in which a test is presented can easily be changed to remove unintended disadvantages for English language learners or for those with disabilities. • Horizontal text. • No construct-irrelevant graphs or pictures; graphics are simple and clear. • Keys and legends at top or right of item. • No time limits.	**0** More than 1 violation	**1** 1 violation	**2** No violations
Simple, clear, and intuitive instructions and procedures • Consistent instructions (e.g., circling correct answer). • Directions allow students to work independently without questions. • Practice or sample items are provided. • Numbered items.	**0** More than 1 violation	**1** 1 violation	**2** No violations
Maximum readability and comprehensibility Plain language and well-constructed sentences should be used for items and directions. Questions should be clearly framed. Verbal and organizational complexity should be minimized. • Simple, clear, common words; no unnecessary words. Technical terms clearly defined. • No typos or spelling errors. • Short sentences. No compound sentences. Noun-pronoun link clear. • Sequenced instructions.	**0** More than 2 violations	**1** 1 or 2 violations	**2** No violations
Maximum legibility Items and instructions should be easily deciphered. This applies to tables, figures, and graphics, as well. Legible tests have high contrast, large font size, and much "white space." • Off-white paper. Black type. At least 50% "white space." • Font is at least 10 point. Graphic text has at least 12-point font. • Standard typeface. Purposeful bolding is ok; otherwise text should generally not be in bold. Unjustified text. • Standard use of upper- and lowercase.	**0** More than 2 violations	**1** 1 or 2 violations	**2** No violations

Source: Adapted from Frey & Allen, 2010.

Precisely defined constructs Performance should not be affected by "construct-irrelevant variance."	**0**	**1**	**2**
• Points awarded for knowledge or performance, not irrelevant tasks (e.g., speed, handwriting, perhaps spelling, and grammar). • Wording for math problems should be simple and clear.	Score appears unrelated to construct	Score appears moderately related to construct	Score appears strongly related to construct
Total Score:			

ASSESSING THE ASSESSMENT

Six of the seven standards for universal test design can be observed directly (sort of) by just looking at a test. A scoring rubric that can be used to assess the extent to which any classroom assessment reflects universal design principles has been developed based on those standards (Frey & Allen, 2010). The rubric is presented as Figure 9.3. You can use this to evaluate how well any assessment (your own or others') follows the guidelines of universal design.

Ms. Clark Believes Variety Is the Spice of Life (Part II)

Stories From the Classroom

Ms. Clark knew she needed to be reminded of all the details of designing a universally "accessible" test, so she turned to her textbook (which she did still have), some old handouts, and some links to online resources from her college days. As she looked over the broad seven standards of universal design, she began asking herself questions to identify things she could do better on her test. Fortunately, it seemed she would not have to make major changes to most of her favorite questions.

She looked through her questions to make sure that irrelevant knowledge or skills were not required. This was fairly easy, because she had been careful in the past to define her assessment objectives to match her important instructional objectives.

The first adjustments she needed were to make sure that her test was technically sound. These were minor changes that she knew could make a big difference for some of her students. She wanted maximum legibility, so the font size was adjusted to 12 point, and the spacing between the questions was also adjusted to ensure more white space on each page. There were also a few questions that required purposeful bolding to make sure the students saw the key parts of the sentence.

After she was satisfied that the test was maximally legible, Ms. Clark focused on the readability of her test. Because her class contained both English Language Learners (ELL) and students with learning disabilities, she verified that her sentences were concise with simple, clear, and common words. Ms. Clark then read through the directions portion to make sure that those were just as readable. She added sample items at the front of her tests to demonstrate what was expected. Finally, Ms. Clark altered the directions slightly, making sure they were consistent throughout the test.

Another issue that Ms. Clark noticed was that one of her charts contained information about cartoon television shows that some of the students may not know. She worried this would confuse or worry some students and affect comprehension. So she removed those pop culture references and also went through the test to make sure words and phrases that are not commonly used across different cultures were changed. Really getting into it now, Ms. Clark asked a translator who worked in the school office to look it over and provide additional suggestions. After that was done, Ms. Clark felt more confident about having nonbiased items.

With the testing completed, Ms. Clark looked over the test scores to see if the results had changed since the first test of the year. Many of the students who Ms. Clark was concerned with had performed better. Even better, the class as a whole performed higher, and very few students had questions during the test. Now her tests would be ready for whatever or whoever came next. And Ms. Clark always enjoyed a little variety in her life.

THINGS TO THINK ABOUT

1. Can a standardized test follow all the principles of universal design? What about a classroom assessment?
2. What elements of universal design would you include in a test meant for a student whose first language is not English?
3. Do you think there is such a thing as a test that is fair for everyone?
4. What are some populations of students to whom classroom assessments are sometimes not fair that have not been mentioned in this chapter?

Looking Back in This Chapter

- Universal test design is derived from a philosophy originally developed in architecture and engineering.
- Teachers often struggle with designing assessments that measure the intended constructs (e.g., knowledge or skill) without scores being affected by construct-irrelevant student characteristics, such as culture, disability, or primary language.
- Assessments can be made more fair, valid, and meaningful by following guidelines for universal design in their development and administration.

ON THE WEB

Three principles of universal design for learning in general
http://www.udlcenter.org/aboutudl/udlguidelines

Universal design of large-scale tests
http://www2.lexcs.org/osep/pdfUniversal_Design_LSA.pdf

Guidelines for teaching linked to rubric for evaluating classroom assessments
http://www.cte.ku.edu/preparing/design/designStrategies.shtml

Universal design and collaborative classrooms
http://www.sst11.org/Files/PDFs/Hines%20114.pdf

STUDENT STUDY SITE

Visit **www.sagepub.com/frey** to access additional study tools including eFlashcards, web quizzes, web resources, additional rubrics, and links to SAGE journal articles.

REFERENCES

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

Bowe, F. G. (2000). *Universal design in education*. Westport, CT: Bergin and Garvey.

Bowe, F. (2005). *Making inclusion work*. Englewood Cliffs, NJ: Prentice Hall.

Brown, P. J. (1999). *Findings of the 1999 plain language field test: Inclusive comprehensive assessment system*. Newark, DE: Delaware Education Research and Development Center, University of Delaware.

Center for Universal Design. (1997). *The principles of universal design*. Raleigh: North Carolina State University.

Dolan, R. P., & Hall, T. E. (2001). Universal design for learning: Implications for large-scale assessment. *IDA Perspectives*, *27*(4), 22–25.

Frey, B. B., & Allen, J. P. (2010, May). *Assessing universal design for classroom testing*. Presented at the Annual Meeting of the American Educational Research Association, Denver, CO.

Gaster, L., & Clark, C. (1995). *A guide to providing alternate formats*. West Columbia, SC: Center for Rehabilitation Technology Services. (ERIC Document No. ED 405689)

Kansas Department of Education. (2010). *Kansas assessment examiner's manual*. Lawrence, KS: Center for Educational Testing and Evaluation.

Ketterlin-Geller, L. R. (2005). Knowing what all students know: Procedures for developing universal design for assessment. *Journal of Technology, Learning, and Assessment*, *4*, 2. Retrieved from http://www.jtla.org

Lombardi, A. R., & Murray, C. (2011). Measuring university faculty attitudes toward disability: Willingness to accommodate and adopt universal design principles. *Journal of Vocational Rehabilitation*, *34*(1), 43–56.

Mcguire, J. M., Scott, S. S., & Shaw, S. F. (2006). Universal design and its applications in educational environments. *Remedial and Special Education*, *27*(3), 166–175.

Messick, S. (1993). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.). Washington, DC: American Council on Education.

Preiser, W. F. E., & Ostroff, E. (Eds.). (2001). *Universal design handbook*. New York, NY: McGraw-Hill.

Rakow, S. J., & Gee, T. C. (1987). Test science, not reading. *Science Teacher*, *54*(2), 28–31.

Rose, D. H., Hall, T. E., & Murray, E. (2008). Accurate for all: Universal design for learning and the assessment of students with learning disabilities. *Perspectives on Language and Literacy*, *34*(4), 23–28.

Salend, S. (2009). Using technology to create and administer accessible tests. *Teaching Exceptional Children*, *41*(3), 40–51.

Thompson, S. J., Johnstone, C. J., Anderson, M. E., & Miller, N. A. (2005). *Considerations for the development and review of universally designed assessments*. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large-scale assessments* (Synthesis Report 44). Minneapolis, MN: National Center on Educational Outcomes.