Research Report

Design Patterns for Improving Accessibility for Test Takers With Disabilities

Eric G. Hansen

Robert J. Mislevy

September 2008

ETS RR-08-49

Design Patterns for Improving Accessibility for Test Takers With Disabilities

Eric G. Hansen
ETS, Princeton, NJ

Robert J. Mislevy
University of Maryland, College Park

September 2008

Abstract

There is a great need to help test designers determine how to make tests that are accessible to individuals with disabilities. This report takes *design patterns*, which were developed at SRI for assessment design, and uses them to clarify issues related to accessibility features for individuals with disabilities—such as low-vision and blindness—taking a test of reading. Design patterns appear useful in clarifying how variable features of a test design need to be matched to disability-related characteristics of test takers in order to ensure accessibility. Giving consideration to accessibility issues during the development and use of design patterns may help improve the validity and fairness of tests, as well as their accessibility for individuals with disabilities.

Key words: Disabilities, assessment design, task design, accommodations, universal design, evidence-centered design

Acknowledgments

The authors acknowledge Malcolm Bauer, Henry Braun, Brent Bridgeman, Linda Cook, Daniel Eignor, Kim Fryer, and Deanna Morgan for their reviews of this document. Any errors are the responsibility of the authors.

Table of Contents

Page

Introduction

There is a great need to help test designers determine how to make tests accessible for individuals with disabilities. Educational accountability systems are increasingly expected to include all students, including individuals with disabilities. Furthermore, there is a moral imperative to ensure that all students, including individuals with disabilities, have access to assessment products and services. Tests need to be designed, developed, and implemented in ways that eliminate accessibility barriers without compromising validity. Yet test design is a process of achieving certain goals under constraints, which requires various tradeoffs. For example, even in cases in which it may be clear how to remove an accessibility barrier, it may not be clear how to do this without undermining the validity of test results.

Design patterns are task-design aides that have been developed through the National Science Foundation-supported project Principled Assessment Designs for Inquiry (PADI), originally for assessing science inquiry skills. Design patterns can help achieve the task-design goals in light of accessibility considerations by providing a way of representing designs that are sensitive to the issues of both validity and accessibility for test takers with disabilities. The design pattern that is the topic of this report concerns a hypothetical reading comprehension test for elementary school students and illustrates how design patterns can help test designers reason through issues related to accessibility features for people with disabilities. This example gives most attention to individuals who have low vision, but the principles used in this analysis appear to be applicable more broadly. The final section summarizes key principles and provides recommendations for further work. This report draws not only upon the work of the PADI project, but also recent accessibility-related extensions to ETS's Evidence Centered Assessment Design (ECD) approach (Hansen, Mislevy, Steinberg, Lee, & Forer, 2005).[1]

Design Patterns

> The term design pattern was coined in the mid-1970s by Christopher Alexander, an architect, who abstracted common design patterns in architecture and formalized a way of describing the patterns in a *pattern language*. A design pattern concerns a problem that occurs repeatedly in our environment, and the core of the solution to that problem—but at a level of generality that the solution can be applied many times without ever being the same in its particulars (Mislevy et al., 2003, p. 8)

The PADI project describes how the concept of design patterns has been applied to assessment design:

> Design patterns lie in the layer in the assessment system called domain modeling. Domain analysis is the activity of identifying the knowledge and skills in a particular subject area to be assessed. Domain modeling specifies the relationships among the knowledge and skills in the area to be assessed. Design patterns are one example of a domain modeling tool. In the case of PADI, the domains of interest are a mix of science content and inquiry processes. The design pattern specifies, in nontechnical terms, the evidence-centered assessment argument and bridges the content expertise and measurement expertise needed to create an operational assessment.
>
> The technical layers of the assessment system are where the details of psychometric models, scoring rubrics or algorithms, presentation of materials, interactivity requirements, and so on, are specified. This technical work can be carried out in accordance with one or more design patterns that lay out the substantive argument of the planned assessment in a way that coordinates the technical details (Mislevy et al., 2003, pp. 4-5)

ECD and Accessibility

ECD was formulated at ETS by Robert Mislevy, Linda Steinberg, and Russell Almond (2003). ECD seeks to make more explicit the evidentiary argument embodied in assessment systems, thereby clarifying assessment design decisions. The essential idea of ECD is to lay out the structures and supporting rationales for the evidentiary argument of an assessment. This involves making explicit the claims (the inferences that one intends to make based on scores), the nature of the evidence that supports those claims, and so on. By making the evidentiary argument more explicit, one makes it more amenable to examination, sharing, and refinement. It also makes it more capable of meeting diverse assessment needs caused by changing technological, social, and legal environments.

Evidence-based approaches have proved useful in law, science (e.g., medicine, natural resource exploration), intelligence analysis, and other fields. Evidence-based approaches rely on principles of logic, reasoning, and probability. In the area of educational measurement, evidence-based approaches may be seen as part of a tradition that pays close attention to validity arguments (Spearman, 1904; Cronbach & Meehl, 1955; Messick, 1989, 1994; Kane, 1992). According to Kane:

> Validity is associated with the interpretation assigned to test scores rather than with the test scores or the test. The interpretation involves an argument leading from the scores to score-based statements or decisions, and the validity of an interpretation depends on the plausibility of this interpretive argument. The interpretive arguments associated with most test-score interpretations *involve multiple inferences and assumptions*. An *explicit recognition of the inferences and assumptions* in the interpretive argument makes it possible to identify the kinds of evidence needed to evaluate the argument. Evidence for the inferences and assumptions in the argument supports the interpretation, and evidence against any part of the argument casts doubt on the interpretation. (Kane, 1992, Abstract, emphasis added)

If, as Kane asserts, “most test-score interpretations involve multiple inferences and assumptions,” then, we would argue, there are an especially large number and variety of inferences and assumptions that need to be made explicit when considering tests administered to subpopulations, such as individuals with disabilities. The accessibility extensions to ECD seek to make more visible the chains of inference and their associated assumptions.

The ECD accessibility work described in this report attempts to apply principles of evidentiary reasoning to handle the complexities of the validity argument associated with accessibility features. The key idea is to lay out the evidentiary structures, what may be termed the validity argument (or what may be termed the validation argument [National Research Council, 2004, p. 104]). An assessment argument can be summarized as comprising (a) a claim about a person possessing at a given level a certain targeted proficiency, (b) the data (e.g., scores) that would likely result if the person possessed, at a certain level, the targeted proficiency, (c) the warrant (or rationale, based on theory and experience) that tells why the person’s level in the targeted proficiency would lead to occurrence of the data, and (d) *alternative explanations* for the person’s high or low scores (i.e., explanations other than the person’s level in the targeted proficiency). The existence of alternative explanations that are both significant and credible might indicate that validity is threatened or being compromised (Messick, 1989). Much of the analysis that is the focus of this report has to do with these alternative explanations, factors that can hinder an assessment from yielding valid inferences. When such alternative explanations are recognized at the earliest stages of test design, then later rework and retrofitting can be avoided.[2] The existence of alternative explanations that are both significant and credible might indicate that validity has been compromised. An example of an alternative explanation for "poor" performance by an individual with a disability is that the individual is not able to receive the test content because there is a mismatch between the test format (e.g., visually displayed text) and the individual's disability (e.g., blindness). An example of an alternative explanation for "good" performance would be that the accommodation eliminates or significantly reduces demand for some aspect of the targeted proficiency. The ECD accessibility effort has focused on building argument structures that might help anticipate and address key details of these alternative explanations particularly as they relate to test takers with disabilities.[3]

Using a variety of methods, the authors have developed models of these arguments that can help reason about a wide variety of situations (diversity in definitions of the construct, task conditions, student characteristics, etc.). Researchers have found Bayes net editing systems to be useful tools in representing complex argument structures produced in ECD domain modeling for subpopulations (Hansen et al., 2003; Hansen & Mislevy, 2004; Hansen et al., 2005; Hansen, Mislevy, & Steinberg, 2008). A Bayes net consists of a set of variables, a graphical structure connecting the variables, and a set of conditional distributions. One then adds evidence to a Bayes net. Adding evidence can take the form of either (a) observing the values of certain variables and then studying the implications for other variables in the network or (b) hypothetically treating certain variables as if their values were known in order to carry out what-if analyses that illuminate implications for other variables in the network. Our work has focused primarily on this latter form of adding evidence. Changes made to these values propagate according to Bayes theorem, yielding updates (posterior values) for each of the other variables. Once such models have been constructed in Bayes net software, it is possible to quickly work through the validity implications of various combinations of test taker profiles, task characteristics, definitions of the targeted proficiency, etc. See other works for information about Bayes nets in assessments (Mislevy, 1994) and more generally (Jensen, 1996). While such models cannot mechanistically make key decisions, they can illuminate the nature of the decisions and help assessment designers think through the sometimes-competing goals of assessment designs.

Universal Design and Related Terms

The ECD accessibility work has often distinguished between two kinds of accessibility features—accommodations and universal design features. Both of these terms are relevant in the discussion of design patterns that follows.

1. *Accommodations*. In the context of educational testing, the term accommodation has been defined as “any action taken in response to a determination that an individual’s disability or level of English language development requires a departure from established protocol” (National Research Council, 2004, p. 1; definition adapted from AERA, APA, NCME, 1999, p. 101). Accommodations are typically broken out into five categories: setting (e.g., separate testing location, individual administration), timing/scheduling (e.g., extended testing time, frequent breaks), presentation (reading aloud by live reader, prerecorded audio, or synthesized speech; font enlargement), response (student dictates answer to scribe, student types instead of writes by hand), and other (use of bilingual word lists, dictionaries).

2. *Universal design features*. In contrast to the idea of an accommodation, which involves a departure from established protocol (National Research Council, 2004, p. 1), a universal design feature is not necessarily a departure from established protocol per se but rather part of a new or refined protocol that seeks to be more inclusive and attentive to individuals’ accessibility needs and preferences.

Another important and related term is *universal design of assessments*, which we use as referring to the design goal of making assessments both accessible and valid for their intended purpose. According to the National Center on Educational Outcomes (NCEO):

> The term “universally designed assessments” refers to assessments that are designed and developed from the beginning to be *accessible and valid* for the widest range of students, including students with disabilities and students with limited English proficiency (National Council on Educational Outcomes, 2004, emphasis added).

Thus, both accommodations and universal design features are strategies for achieving or at least moving toward the ideal of universally designed assessments, that is, assessments that are both accessible and valid. This focus on both accessibility and validity is consistent with a growing awareness of the importance of safeguarding against unacceptable compromises to validity in the course of applying accessibility principles (Heath & Hansen, 2002; Thompson, Johnstone, & Thurlow, 2002; Thompson & Thurlow, 2002; Hansen et al., 2003; Hansen, Forer, & Lee, 2004).

Other Resources

There are other useful knowledge resources for reasoning through the design issues related to test takers with disabilities. For example, there is an emerging consensus about the kinds of features that will help lower accessibility barriers.

- The Web Content Accessibility Guidelines of the World Wide Web Consortium (W3C) guides content authors in developing Web content that lowers accessibility barriers (Chisolm, Vanderheiden, & Jacobs, 1999). The W3C's User Agent Accessibility Guidelines help developers of Web browsers and media players develop accessible software (Jacobs, Gunderson, & Hansen, 2002).
- The Section 508 standards set accessibility standards for information technology procured by the U.S. government (Architectural and Transportation Barriers Compliance Board, 2000).[4]
- The National Center for Accessible Media (NCAM), which is part of WGBH in Boston, has provided guidelines for accessible math and science software and Web sites (Rothberg & Wlodkowski, 2000).
- The IMS Global Learning Consortium accessibility group has developed two relevant specifications. First is the Accessibility for Learner Information Package version 1.0 (ACCLIP) specification (IMS Global Learning Consortium, 2003), which allows the accessibility needs and preferences of individuals to be stored. The set of ACCLIP need/preference settings for a given student constitutes a student profile. The second specification is the AccessForAll Meta-data 1.0 (ACCMD) specification, which specifies mark-up that will allow the accessibility-related significance of the content to be recognized (IMS Global Learning Consortium, 2004). With the student profile provided by ACCLIP and the resource profile provided by ACCMD, a system has the beginnings of a basis for matching accessible resources to people who need or desire them, thus helping automate the delivery of accessible online content. A white paper

on accessible online learning systems (IMS Global Learning Consortium, 2002) developed by the accessibility group includes a section on testing and assessment (Heath & Hansen, 2002).

- The NCEO tracks state and national practices and trends in accommodations. Based on NCEO work by Thurlow, House, Boys, Scott, and Ysseldyke (2000) that surveyed states' policies regarding accommodations and participation in state assessments, Sheinker, Barton, and Lewis (2004) have placed testing accommodations into three categories, based on whether they alter the interpretation of scores. It seems reasonable to focus intensive use of design patterns or other domain modeling tools on the impact of categories of accommodation that *may* (category 2) or *are likely* (category 3) to alter test-score interpretations as opposed to a category that is *not expected* (category 1) to alter such interpretations.

Attributes of Design Patterns

Users of design patterns—whether they are test designers, testing program directors, teachers, or others—can use such patterns as tools to think through key design elements and their relationships with each other. Table 1 describes the attributes of a PADI design patterns as described by Mislevy et al. (2003).

Some key attributes. Let us consider four key attributes of design patterns: (a) focal knowledge, skills, and other attributes (focal KSAs); (b) additional KSAs; (c) characteristic features; and (d) variable features.

1. *Focal KSAs.* According to Mislevy et al. (2003, p. 36), focal KSAs consist of "the primary knowledge/skills/attributes of students that are addressed" by the assessment.[5] Ordinarily, comparability of scores between individuals with and without disabilities is important, which suggests that one should seek evidence about the same set of focal KSAs, regardless of whether the test taker has a disability or not.

2. *Additional KSAs*[6] The test designer needs to also attend to additional KSAs, the "other knowledge/skills/attributes that may be required in a task" (Mislevy et al., 2003, p. 36). For tests of academic subjects, the abilities to see and hear are typically additional KSAs. On the other hand, for assessments of sight and hearing, respectively, sight and hearing are likely to be defined as focal KSAs. Notice that

there are many disabilities that involve impairments of sight, hearing, or both (e.g., blind, low vision, color-blind, deaf, hard of hearing, deaf-blind). Deficits in such additional KSAs can cause unduly low scores among test takers with disabilities.

3. *Characteristic features.* Characteristic features of the assessment consist of the "features that must be present in a situation in order to evoke the desired evidence" about the focal KSAs (Mislevy et al., 2003, p. 37). For example, if one is assessing reading comprehension proficiency, then typically an important characteristic feature is "consistent availability of the reading passage while the items are being answered" as opposed to only being able to view the passage once, then having it disappear, which would place higher demand on memory than is likely to be appropriate.[7]

4. *Variable features***.** Variable features are described as features that "can be varied to shift the difficulty or focus of tasks" (Mislevy et al., 2003, p. 37). Variable features have a particularly significant role with respect to test takers with disabilities and other subpopulations (e.g., speakers of minority languages). Much of our attention in this report will be on manipulating variable features to reduce or eliminate demands for additional KSAs in which there may be a deficit while making sure (to the extent possible) that demands for focal KSAs have not been changed.[8]

Similarities and differences in roles of key attributes. Most of the key attributes of the design pattern have essentially the same values and functions for candidates with disabilities as for those without. For example, the test designer must carefully distinguish between focal KSAs and additional KSAs, and this basic distinction holds regardless of whether the test taker in question has a disability. Similarly, the test designer must anticipate the way in which characteristic features and variable features drive demand for the test taker's focal KSAs and additional KSAs so that the test properly measures, to the extent feasible, the targeted proficiency (composed of focal KSAs) rather than the additional KSAs (including those that characterize disabilities).

However, as suggested earlier, there are also some differences in the four key attributes relative to test takers with and without disabilities. For tests takers without any disability, we

Table 1

Attributes of a PADI Assessment Design Pattern

Attribute	Definition
Title	A short name for referring to the design pattern.
Summary	Overview of the kinds of assessment situations students encounter in this design pattern and what one wants to know whether they can do based on their knowledge, skills, and abilities.
Rationale	Why the topic of the design pattern is an important aspect of scientific inquiry.
Focal KSAs	Primary knowledge/skills/abilities of students that one wants to know about.
Additional KSAs	Other knowledge/skills/abilities that may be required.
Potential observations	Some possible things one could see students doing that would give evidence about the KSAs.
Potential work products	Different modes or formats in which students might produce the evidence.
Potential rubrics	Links to scoring rubrics that might be useful.
Characteristic features	Kinds of situations that are likely to evoke the desired evidence.
Variable features	Kinds of features that can be varied in order to shift the difficulty or focus of tasks.
I am a kind of...	Links to other design patterns that this one is a special case of.
These are kinds of me	Links to other design patterns that are special cases of this one.
I am part of...	Links to other design patterns that this one is a component or step of.
These are parts of me	Links to other design patterns that are components or steps of this one.
Educational standards	Links to educational standards.
Templates (task/evidence shells)	Links to templates that use this design pattern.
Exemplar tasks	Links to sample assessment tasks that are instances of this design pattern.
Online resources	Links to online materials that illustrate or give backing for this design pattern.
References	Pointers to research and other literature that illustrate or give backing for this design pattern.
Miscellaneous associations	Other relevant information.

Note. From *Design Patterns for Assessing Science Inquiry* (p. 29), by R. J. Mislevy, L. Hamel, R. Fried, T. Gaffney, G. Haertel, A. Hafter, R. Murphy, et al. 2003, Menlo Park, CA: SRI International. Copyright SRI International. Adapted with permission.

typically establish a set of default (or standardized) characteristic features and settings of variable features. We typically do this based on the assumption or knowledge of these features being appropriate given the states of additional KSAs that characterize nondisabled test takers. For example, the characteristic feature of having the proctor speak the test instructions is based on a belief that nondisabled test takers will be able to hear those spoken directions. On the other hand, we recognize that a deficit in a particular additional KSA may cause an unnecessary hurdle or barrier for a test taker with a disability (such as deafness), which causes us to want to change from the default features.[9] For example, if the test taker is deaf, we may present the test directions using sign language (instead of having them spoken), thereby eliminating demand for an additional KSA in which the test taker has a disability (hearing) and relying, instead, on an additional KSA in which the test taker has no disability (receptive sign language).

Our goal in providing accessibility features (accommodations and universal design features) for people with disabilities is generally to remove unfair disadvantages while at the same time addressing the possibility of unfair advantages for the person receiving the accommodation. As Robert Linn has said: “The purpose of an accommodation is to remove disadvantages due to disabilities that are irrelevant to the construct the test is intended to measure without giving unfair advantage to those being accommodated” (Linn, 2002, p. 36).

The first step is to remove unfair disadvantages by ensuring that any deficits in additional KSAs are not the cause of poor performance on the test. For example, by changing the value of the font size variable from 12 point to a larger size, we eliminate the need for excellent eyesight and require only that the individual be partially sighted (i.e., have low vision). Another example would be a person who is blind to whom we may present test content in Braille rather than as visually displayed text. This accommodation gains its effectiveness for this individual by relying on additional KSAs in which there is no deficit; that is, the individual relies on the sense of touch and knowledge of Braille codes instead of the sense of sight, in which there is a deficit. If we can make these changes without giving unfair advantages to the person receiving the accommodation by changing, especially decreasing, demand for the focal KSAs, then there is a good chance that we have provided an appropriate accommodation.[10]

The design pattern that is the focus of this report focuses mostly on how one removes unfair disadvantages. Additional discussion at the end of the report gives more detailed consideration of the challenge of recognizing unfair advantages and determining how to deal with those challenges.

The Design Pattern

Table 2 illustrates a design pattern using the example of a reading comprehension test for elementary school students. This example attempts to reflect an awareness of test takers with visual disabilities—particularly low vision—as well as individuals who are nondisabled. This example assumes the viability and validity (for appropriate nondisabled native speakers of English) of administering the test via hard-copy print using single-selection multiple-choice items. While the major focus of Table 1 is on nondisabled and low-vision test takers, it also contains some notes and comments pertaining to other student profiles.

Column 1 of the table lists the standard attributes for design patterns (see Table 1). Column 2 lists values of those attributes for test takers who are nondisabled, and column 3 lists values of those attributes for test takers with *low vision*.[11] Finally, column 4 provides additional comments.

While the design pattern table is intended to be largely self-explanatory and nontechnical, let us consider a couple of important rows in the table—*Additional KSA(s)* and *Variable features*. These two rows are closely related to each other with respect to test takers with disabilities. Perhaps most important for this discussion, variable features play an essential role in *reducing or eliminating the demands for additional KSAs* that are *not satisfied by the test taker's abilities*. In doing so, variable features can remove accessibility barriers.

Let us consider how Additional KSAs and Variable features are related with respect to the sense of sight. (Recall that the disability aspect of this table is focused largely on the issue of low vision.)

Consider first the Additional KSA(s) row. Note that in column 2 (*Values for test takers who are nondisabled*), the first KSA listed is the sense of sight (*See*) and the value (for nondisabled test takers) is *yes* (i.e., see = yes), meaning that the test design asserts that the nondisabled test taker must be able to see in order to perform well on the test when administered in default conditions.

Default conditions are specified at a number of locations in the table (e.g., characteristic features, variable features, etc.); however, for the feature that induces the requirement (or demand) for the sense of sight—*font size*—the values area is specified in the *Variable features*

Table 2

Design Pattern for Reading Comprehension Test

1. Attribute	2. Values for test takers who are nondisabled	3. Values for test takers with low vision	4. Comments
Title	Reading comprehension (RC)—Elementary school level	(Same as for test takers who are nondisabled)	
Summary	In this design pattern, students are presented with a passage to read. Can they understand the written content passage, identify its main idea, and make inferences on the basis of what they have read?	(Same as for test takers who are nondisabled)	
Rationale	Studies show that reading comprehension skills differentiate more successful students from less successful students.	(Same as for test takers who are nondisabled)	
Focal KSAs	RC. In this design pattern, let us assume that RC has two possible values: good and poor.	(Same as for test takers who are nondisabled)	As will be discussed later, the greater the diversity of disabilities being addressed, the greater the need to clarify distinctions between focal KSAs and additional KSAs.

Table continues

Table 2 (continued)

1. Attribute	2. Values for test takers who are nondisabled	3. Values for test takers with low vision	4. Comments
Additional KSAs	Important additional KSAs might include KSAs such as *see* (sense of sight); know the topics (of the passages); know vocabulary in English; work quickly; physically pencil-in the answers on an answer sheet; hear and follow spoken directions. Generally, one should keep requirements for additional KSAs as low as is feasible, which increases the probability that an individual's values of these KSAs will be sufficient to perform well. Below are some of the specific values of these additional KSAs that are important for a nondisabled test taker. 1. "See = yes." The individual must be able to see well. Let us assume that this means that a person can receive text in 12-point font. 2. "Know the topics (of the passages) = low to moderate." It is important to have some familiarity with the topics. 3. "Know vocabulary in English = low to moderate." 4. "Work quickly = low." We assume here that the test does not include the ability to work quickly as part of the focal KSAs. However, because the time allowed is not unlimited and experience has shown that a small percentage of individuals do not finish the test in the default allotted time (40 minutes), the test has at least a small degree of unintentional (construct-irrelevant) speededness, thus requiring a *low* (rather than "*none*" or "*high*" degree of the ability to work quickly). 5. "Physically pencil-in answers on an answer sheet = yes." One must be able to physically pencil in the bubbles on an answer sheet.6. "Hear and follow spoken direction = yes." One must be able to hear and follow spoken directions.	For the individuals with low vision, the set of additional KSAs is essentially the same as for nondisabled individuals, except for the following. 1. "See = partial." In this example, someone with "see = partial" is partially sighted. By this we mean that they have low vision and that they require a large font (greater than 12 point) in order to use testing materials.[12]	Note that for any additional KSA, there is potential for a test taker (particularly an individual with a disability or a non-native speaker of English) to lack the necessary levels to do well. For example, for a person with a physical disability that impairs the use of hands, filling in the answer bubbles may be impossible. In this example, we are focusing primarily on low vision.

Table continues

Table 2 (continued)

1. Attribute	2. Values for test takers who are nondisabled	3. Values for test takers with low vision	4. Comments
Potential observations	Scores on individual items or parts of items might be indicative of any of the following: 1. deduction of the main idea of a passage 2. identification of the referent of anaphora 3. drawing conclusion from propositions in the text Ultimately, the observations might be reduced to dichotomous (e.g., correct/incorrect) or more fine-grained scores.	(Same as for test takers who are nondisabled)	The score might be generated automatically from a mechanical answer sheet scanner or from a human scorer.
Potential work products	There are two major categories of work product: 1. Multiple choice. For example, marks on a scannable answer sheet indicate the test taker's selection of an option on single-selection multiple-choice test format. OR 2. Constructed response. For example, expressing the words of the response, as recorded on paper or typed into a computer.	This is essentially the same as for nondisabled test takers. However, if the low vision is accompanied by an additional disability, such as a physical disability, then it may be necessary to allow the test taker to dictate his or her answers to a scribe or allow the test taker to mark the answers directly onto the paper test.	

Table continues

Table 2 (continued)

1. Attribute	2. Values for test takers who are nondisabled	3. Values for test takers with low vision	4. Comments

Potential rubrics	Use the rule:
	“IF (Answer Key = Answer Given By Student)
	THEN (Item Score = “Correct”),
	ELSE (Item Score = “Incorrect”)”
		(Same as for test takers who are nondisabled)	If the items were constructed response items, the more complex scoring rules or rubrics would apply.
			In an operational test, scores from multiple tasks would be accumulated to produce a higher-level score, such as a section score or a test score.

Characteristic features	1. The passage remains available while the test taker attempts to answer the item (as opposed to being able to only read it once). 2. Items all have “high” reading comprehension demand.[13] 3. Three sets for the entire test. 4. Each set has a passage of about 200 words and four items. 5. Each item has a stem and five options.	(Same as for test takers who are nondisabled)	Comprehension demands of items should be high since, for this example, the tendency to answer high-RC-demand items distinguishes between individuals with good versus poor RC ability.[14]

Table continues

Table 2 (continued)

1. Attribute	2. Values for test takers who are nondisabled	3. Values for test takers with low vision	4. Comments
Variable features	Some variable features are intended to modify emphasis or to vary item difficulty, while others could be intended to provide access for diverse learners, such as individuals with disabilities or those whose first language is not English. For the purpose of illustrating the idea of manipulating variable features so that the test taker’s values of additional KSAs exceed the threshold needed for good performance, there are two key features worth considering: (a) font size and (b) testing time. For a person who is nondisabled, the default values of font size (regular, i.e., 12 point) and testing time (i.e., 40 minutes) is sufficient. Other variable features might include those that govern the emphasis on potential observations such as (a) deduction of the main idea of a passage, (b) identification of the referent of anaphora, and (c) drawing conclusions from propositions in the text (Mislevy, 1994; Sheehan & Ginther, 2001).	For the person with low vision, it is necessary that font size = large (i.e., larger than 12 point). Large font size should apply both to test items and to other test materials.[15] For individuals who have low vision and are receiving content with a large font size, let us suppose that they generally receive some extra testing time (e.g., an extra 10 minutes or 1.25 times the default time of 40 minutes)[16]	Note that other variable features could come into play for test takers with disabilities, including: 1. Mode for presenting test content (visual text, readaloud, Braille). 2. Mode for presenting directions (sign language, etc.). 3. Response mode (write on scannable answer sheet, use human scribe, type into a computer).

Note. The names of the attributes have been brought into conformance to the names of the first 10 attributes in Table 1 in this document. The content of the columns is closely adapted from Hansen (2002).

row in column 2. We see that in default conditions, the reading comprehension test is displayed in hard copy in regular-sized (i.e., 12-point) font.

We thus can see the linkage implied by this information. Specifically, the test design asserts that the sense of sight for a person for whom see = yes (and without any disability) should be able to satisfy the demand for sight imposed by the use of regular-sized font. In other words, for that nondisabled person, the regular-sized font imposes no accessibility barrier.

Now let us consider test takers with low vision. We begin examining additional KSAs in column 3 (*Values for test takers with low vision*) and we see the value for sight for low vision test takers is partial, that is, see = partial. Now examine variable features in column 3, where we see that the variable *font size* is set to large. Thus, the test design asserts that the test taker who has low vision (see = partial) should be able to satisfy the demand for sight imposed by large font size (font size = large). In other words, for the person with low vision, a large-sized font should impose no accessibility barrier.

In summary, when individuals with disabilities face accessibility barriers due to demands for additional KSAs that their own abilities are not able to satisfy, an accessibility feature might be implemented for one or more variable features (e.g., font size). Alterations in these variable features may reduce or eliminate demands for additional KSAs in which there is a deficit (due to a disability). In the example we just examined, a person with low vision is unable to satisfy the demand for sight imposed by regular-sized font; therefore, an alteration in the font size (from regular to large) reduces (but does not eliminate) the demand for sight so that their partial sight is enough to satisfy the demand, thereby eliminating the accessibility barrier.

It may appear that this level of detail in analysis is excessive, particularly for an accessibility feature as basic as large font, yet the basic logic seems applicable and extensible for more complex cases, such as the readaloud feature (having test content read aloud via a live reader, synthesized speech, or prerecorded audio) on reading and related tests. (This is discussed later in this report.).

In general, the test design should help ensure that each test taker can satisfy the demands for additional KSAs. If the design further ensures that each test taker is presented with the same, appropriate demand for focal KSAs, then the scores produced are likely to be valid.[17] That is, those who possess the targeted proficiency will tend to perform well and those who do not possess the targeted proficiency will tend to perform poorly. Yet if there are unsatisfied demands for additional KSAs then there is arguably an accessibility barrier and scores will tend to be invalid.

Discussion

Our basic strategy for providing accessibility features is to manipulate variable features of an assessment design to reduce demand for additional KSAs in which there is a deficit and, at the same time, maintain demand for focal KSAs.

Using this strategy, we have reason to believe that providing a larger font size for a person with low vision could be an appropriate accessibility feature. It must be noted that this judgment is based on a number of assumptions. For example, we assume that we know that the candidate actually has low vision; that they have already used and benefited from this accommodation in instructional and/or testing settings; that they have had the opportunity to become familiar with our specific way of implementing this feature with sample items; and that it is technically and practically feasible to implement the large fonts.

Variable features deserve considerable attention from test designers since, of the four key attributes that we are examining (focal KSAs, additional KSAs, characteristic features, and variable features), it is the one over which test designers generally have the most control. The decision about what constitutes a focal KSA (as opposed to additional KSAs) should, in principle, be made once, by the test designer (perhaps in conjunction with a panel of experts), and then remain largely unchanged thereafter.[18] Characteristic features are, by definition, stable and not subject to manipulation.[19] In contrast, variable features are, by definition, variable and are at the disposal of test designers and others to lower accessibility barriers while safeguarding validity.[20] Generally, once a KSA has been determined to be additional rather than focal, then demand for that additional KSA should be made as low as is feasible. Reducing demand for additional KSAs does not—of itself—reduce assessment validity, provided it does not also cause a reduction in demand for the focal KSAs. Indeed, evidence that demands for the focal KSAs have not been reduced is a defense against the assertion that a given accessibility feature has reduced the rigor of the test.

This strategy is conceptually simple and is often relatively easy to carry out. For example, for a test of reading comprehension, it seems clear that using a larger font size should not appreciably change demand for reading comprehension ability, so that font enlargement is likely an appropriate accessibility feature for someone with low vision. On the other hand, if one were assessing visual acuity, it seems clear that enlarging the font size may not be appropriate if it would reduce demand for visual acuity or its constituent KSAs.

However, there can be subtleties that make the situation more complex. Ordinarily, we might assume that a person with unimpaired vision will perform equally well with regular font *or* large font. However, empirical research on font size may complicate that picture with findings that say that large font can be a disadvantage for some disabled and nondisabled individuals, perhaps because it tends to involve more scrolling or page turning. Depending on the nature of the accessibility decision, it may or may not be important to attempt to address such subtleties within a design pattern. Even if such fine points of research knowledge do not become explicit parts of a design pattern, they can play a role in guiding the wise use of the design pattern.

A More Difficult Case

Some cases can be considerably more difficult. For example, in some cases it is difficult to recognize whether an accessibility feature would reduce demand for focal KSAs and thereby compromise validity. Consider the case of a person who is blind and requests an accommodation of having the reading comprehension test read aloud via a live reader, synthesized speech, or prerecorded audio. This accommodation may be represented as modifying a variable feature called *presentation mode* from *visually displayed text* to *readaloud*.[21] This accommodation appears to overcome an accessibility barrier by allowing the person who is blind to receive the test content. Specifically, relying on the additional KSA of hearing, which has no deficit, the accommodation eliminates demand for the additional KSA of sight, which has a deficit, thus overcoming the barrier to reception. At first glance, this accommodated situation would be one in which the test taker could demonstrate his or her true reading comprehension ability.

Yet, there is a potential for failing to recognize the possibility for reducing demand for the focal KSA of reading comprehension. Note that a readaloud accommodation generally involves speaking whole words rather than speaking one letter at a time. The readaloud mode of content delivery thus essentially eliminates demand for decoding (the ability to form words from letters). If decoding is part of the targeted proficiency of reading comprehension, then providing the readaloud accommodation may confer an unfair advantage on the person receiving the accommodation. Specifically, such an accommodation would allow a person with poor reading comprehension due to poor decoding ability to perform better than they should, thus yielding an unfair advantage to the person receiving the accommodation.[22] In other words, where the targeted proficiency of reading comprehension includes decoding, then the readaloud presentation mode would tend to compromise validity. Specifically, the person who has a poor (i.e., low) level in the targeted proficiency and a good (high) in their measured proficiency (or performance) may be termed as having a *false-positive* outcome. On the other hand, if reading comprehension is defined as consisting of comprehension by itself and decoding is strictly an additional KSA, then the same test content, same test taker, and same accessibility feature yields a valid outcome. This valid outcome may be termed *true-positive* since the person has both a good (i.e., high) level in both his or her targeted proficiency and their measured proficiency or performance. Under this latter definition of the targeted proficiency, the measured focal KSA accurately reflects the intended focal KSA, which is consistent with the goals of validity and fairness.

This example points out the importance of a clear definition of the construct (targeted proficiency) and the impact of that definition of the validity of score interpretations. The definition of the targeted proficiency is decided or specified, rather than being estimated, calculated, or discovered. The definition of the targeted proficiency may be informed by empirical as well as theoretical considerations, but in the end, it is a decision. This definition, however arrived at, becomes the foundation for judging test validity and appropriateness of accessibility features. Empirical findings, such as those concerning the factor structure or test-score boost (Bielinski, Thurlow, Ysseldyke, Freidebach, & Freidebach, 2001; Phillips, 1994) in observed scores, can inform judgments about test validity and appropriateness of accommodations. Yet without a clear definition of the intent of measurement (the construct to be measured), those empirical findings are of limited value relative to the goal of ensuring validity.

Making Decisions About the Use of Accessibility Features

Even if one has recognized a significant probability of an unfair advantage or false-positive for the person using the accessibility feature, there can be additional issues in determining whether or not to actually allow the use of the feature.

For example, part of the challenge of making an accommodation decision may come from a lack of consensus or clarity about how considerations beyond those that are strictly (or narrowly) validity-related should affect accommodation decisions. For example, authors of a recent work on students with disabilities and standards-based educational reform stated: “Accommodations should be offered during large-scale assessments for only two purposes: (1) to facilitate participation by students with disabilities and (2) to increase the validity of scores" (McDonnell, McLaughlin, & Morison, 1997, p. 204).[23] In other words, accommodations (in this view of large-scale assessments) have a role not only in ensuring or increasing validity but also in facilitating participation or inclusion. Having multiple purposes for accommodations can make it difficult to determine what constitutes an appropriate accommodation. For example, while there are undoubtedly many instances in which inclusion can be increased without compromising validity, some potentially serious compromises can arise and one must be able to decide what kinds and levels of inclusion warrant what kinds and degrees of compromise to validity.

For example, suppose one determines that providing a readaloud accommodation for the reading comprehension test will likely result in some compromise to validity. Following are some issues to consider in determining what action to take. What are specific inferences that one wishes to make based on the scores?[24] For example, does one wish to generalize to a situation in which the readaloud accommodation will be available or unavailable to the test taker? Is decoding a truly critical or only moderately important part of the targeted proficiency? What are the actual decoding (and comprehension) demands (or requirements) of the tasks? Could the individual participate by receiving the test content via Braille instead of readaloud? Which is preferable—to obtain somewhat compromised scores via readaloud or to obtain highly compromised scores (or perhaps no scores at all) using default conditions (visually displayed text)? Is it technologically and logistically feasible to deliver a readaloud accommodation? Are there legal or policy considerations that should affect the decision? (Phillips, 2002). Judgments about such issues will inform decisions whether to actually allow the readaloud accommodation.

For most accessibility features, an important issue is whether the test taker's knowledge of the format is sufficient to avoid being a barrier to valid measurement. Thus, "know how to use the test format" could be an important KSA (typically an additional KSA) that the test taker must be able to satisfy. Ensuring that the test taker has adequate knowledge of the format can be facilitated by rules that allow only accommodations that the test taker has already used earlier in instructional settings. The availability of practice and familiarization materials may also be important; indeed, practice and familiarization materials may be important for accommodations (which require prior approval), but may also be important for universal design features to help test takers avoid making unwise use of accessibility features. For example, if the readaloud feature were made available to all test takers without prior approval, it may be important to acquaint all test takers with how to turn the feature on and off as well as to warn the test taker, because having the feature on may prove to be merely a distraction.

A few accessibility features may require no such familiarization steps. Consider, for example, the action of removing unnecessary linguistic complexity from a test. Such a linguistically modified test could be used by all test takers without special approval and would therefore be considered a universal design feature. Such a feature would benefit individuals with disabilities that affect language acquisition and use, such as deafness and certain learning disabilities, as well as benefiting English language learners. Such a feature is pervasive as well as passive in the sense of requiring no special action or choice on the part of the test taker.

Conclusion

Design patterns can play a valuable role in elucidating the core validity issues that relate to the use of accessibility features such as testing accommodations and universal design features. Of particular note is the relationship between design pattern variable features (test features that can be varied) and additional KSAs (knowledge, skills, and other attributes that are not part of what one intends to measure but which may be required in order to perform well on the test). The distinction between the two is foundational in reasoning about the validity of accommodations: Variable features are aspects of tasks, under the control of the test designer, which can induce requirements for additional KSAs. When the additional KSAs a particular task feature evokes is problematic to an examinee, poor performance for a construct-irrelevant reason results. The ideal would be to structure each examinee’s assessment using only variable features that impose demands on additional KSAs the examinee is known to possess, plus the focal KSAs that are the object of measurement.

This report specifically examined the variable feature of font size and related it to the additional KSA of sight (*see*) on a test of reading that is displayed visually. The basic structure of design patterns begins to draw a distinction between focal KSAs and additional KSAs. For some purposes, test designers may wish to extend that reasoning to make the distinction between the intent of measurement (the targeted proficiency, composed on focal KSAs) and what is actually measured (see Hansen et al., 2005).

Design patterns further invite consideration of the specific changes (such as those involving accommodations or universal design features) to variable features of the task situation so that good measurement will be enabled. In addition, dealing with basic issues common to virtually all testing accommodations, design patterns can also help test designers begin to think through the nuances and complexities of some of the more difficult cases. Both the PADI design pattern work and the ECD accessibility work have a common objective in supporting quality and effectiveness in assessment design. While neither design patterns nor ECD can automatically solve some of the most critical issues involving test takers with disabilities, they do seem capable of helping test designers reconcile many important assessment design constraints. It is important that the concepts and practices of such design approaches be applied for the benefit of test takers with disabilities. Existing approaches for test takers with disabilities may suffice for testing programs that are very stable and have evolved over time to respond to commonly encountered challenges and problems. Yet, where testing programs are new, in transition, or facing serious challenge, then there is a need for frameworks that are more flexible. Design concepts such as those used in the PADI project or ECD may play a valuable role in such circumstances, in capturing the rationale behind accommodations practices, and in making it public and available to guide design of new assessments.

References

Aguirre-Muñoz, Z., & Baker, E. (1997). *Improving the equity and validity of assessment-based information systems* (CSE Technical Report 462). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

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

Architectural and Transportation Barriers Compliance Board. (2000). *Electronic and information technology accessibility standards (Section 508)*. Retrieved January 25, 2008 from http://www.section508.gov/index.cfm?FuseAction=Content&ID=12

Bielinski, J., Thurlow, M., Ysseldyke, J., Freidebach, J., & Freidebach, M. (2001). *Read-aloud accommodations: Effects on multiple-choice reading and math items* (Technical Rep. No. 31). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Chisolm, W., Vanderheiden, G., & Jacobs, I. (Eds.). (1999). *Web content accessibility guidelines* (W3C recommendation). Retrieved May 5, 1999, from http://www.w3.org/TR/WAI-WEBCONTENT

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281-302.

Elliott, S. N., & Roach, A. T. (2002, April 3). *The impact of providing testing accommodations to students with disabilities.* Paper presented at the annual meeting of the American Educational Research Association. Retrieved May 27, 2003, from http://www.wcer.wisc.edu/testacc/Publications/aera2002.doc

Haertel, E. H., & Linn, R. L. (1996). Comparability. In G. W. Phillips (Ed.), *Technical issues in large-scale performance assessment* (pp. 59-78). Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement.

Hansen, E. G. (2002). *A design pattern for a reading comprehension test*. Unpublished manuscript.

Hansen, E. G., Forer, D. C., & Lee, M. J. (2004). *Toward accessible computer-based tests: Prototypes for visual and other disabilities* (TOEFL Research Rep. No. RR-78). Princeton, NJ: ETS.

Hansen, E. G., & Mislevy, R. J. (2004, April 13). *Toward a unified validity framework for ensuring access to assessments by individuals with disabilities and English language learners*. Paper presented at the annual meeting of the National Council on Measurement in Education (NCME), San Diego, California.

Hansen, E. G., & Mislevy, R. J. (2005). Accessibility of computer-based testing for individuals with disabilities and English language learners within a validity framework. In M. Hricko & S. Howell (Eds.), *Online assessment and measurement: Foundation, challenges, and issues*. Hershey, PA: Idea Group Publishing, Inc.

Hansen, E. G., Mislevy, R. J., & Steinberg, L. S. (2003). Evidence-centered assessment design and individuals with disabilities. In E. G. Hansen (Organizer), *Assessment design and diverse learners: Evidentiary issues in disability, language, and non-uniform testing conditions.* Symposium presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Hansen, E. G., Mislevy, R. J., & Steinberg, L. S. (2008). *Evidence-centered assessment design for reasoning about testing accommodations in NAEP reading and mathematics* (ETS Research Rep. No. RR-08-38). Princeton, NJ: ETS.

Hansen, E. G., Mislevy, R. J., Steinberg, L. S., Lee, M. J., & Forer, D. C. (2005). Accessibility of tests for individuals with disabilities within a validity framework. *System: An International Journal of Educational Technology and Applied Linguistics, 33*(1), 107-133.

Hansen, E. G., Mislevy, R. J., & Steinberg, L. S. (2003). Evidence-centered assessment design and individuals with disabilities. In E. G. Hansen (Organizer), *Assessment design and diverse learners: Evidentiary issues in disability, language, and non-uniform testing conditions.* Symposium presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Heath, A., & Hansen, E. G. (2002). *Guidelines for testing and assessment (Section 9)*. IMS Guidelines for Developing Accessible Learning Applications. IMS Consortium. Retrieved July 18, 2002, from www.imsproject.org/accessibility/accv1p0/imsacc_guidev1p0.html

IMS Global Learning Consortium. (2002). *IMS guidelines for developing accessible learning applications*. Retrieved September 9, 2002, from http://imsproject.org/accessibility/accv1p0/imsacc_guidev1p0.html

IMS Global Learning Consortium. (2004). *IMS AccessForAll meta-data information model (Version 1.0).* Retrieved September 25, 2008, from http://www.imsglobal.org/accessibility/accmdv1p0/imsaccmd_infov1p0.html

IMS Global Learning Consortium. (2003). *IMS Learner Information Package Accessibility for LIP Information Model (Version 1.0).* Retrieved September 25, 2008, from http://www.imsglobal.org/accessibility/acclipv1p0/imsacclip_infov1p0.html

Jacobs, I., Gunderson, J., & Hansen, E. (Eds.). (2002). *User agent accessibility guidelines 1.0.*. Retrieved September 9, 2002, from http://www.w3.org/TR/UAAG10/

Jensen, F. V. (1996). *An introduction to Bayesian networks.* New York: Springer-Verlag.

Kane, M. (1992). An argument-based approach to validation. *Psychological Bulletin, 112,* 527-535.

Linn, R. L. (2002). Validation of the uses and interpretations of results of state assessment and accountability systems. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students* (pp. 109-148). Mahwah, NJ: Lawrence Erlbaum.

McDonnell, L. M., McLaughlin, M., & Morison, P. (Eds.). (1997). *Educating one and all: Students with disabilities and standards based reform.* Washington, DC: National Academy Press.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Education Researcher, 32*(2), 13-23.

Mislevy, R. J. (1994). Evidence and inference in educational measurement. *Psychometrika, 59,* 439-483.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1,* 3-67.

Mislevy, R. J., Hamel, L., Fried, R., Gaffney, T., Haertel, G., Hafter, A., et al. (2003). *Design patterns for assessing science inquiry* (Principled Assessment Designs for Inquiry [PADI] Technical Rep. No. 1). Menlo Park, CA: SRI International.

National Council on Educational Outcomes. (2004). *Special topic area: Universally designed assessments.* Retrieved August 24, 2004, from http://education.umn.edu/NCEO/TopicAreas/UnivDesign/UnivDesign_topic.htm

National Research Council. (2004). Keeping score for all: The effects of inclusion and accommodation policies on large-scale educational assessment. In J.A. Koenig & L.F. Bachman (Eds.), *Committee on participation of English language learners and students with disabilities in NAEP and other large-scale assessments*. Washington, DC: National Academy of Sciences.

Phillips, S. E. (1994). High stakes testing accommodations: Validity vs. disabled rights. *Applied Measurement in Education, 7*(2), 93-120.

Phillips, S. E. (2002). Legal issues affecting special populations in large-scale testing programs. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students* (pp. 109-148). Mahwah, NJ: Lawrence Erlbaum.

Rothberg, M., & Wlodkowski, T. (2000). *Making educational software accessible: Design guidelines including math and science solutions* (CD-ROM Access Project). Boston: WGBH Educational Foundation.

Sheehan, K. M, & Ginther, A. (2001, April). *What do multiple choice verbal reasoning items really measure? An analysis of the cognitive skills underlying performance on TOEFL reading comprehension items*. Paper presented at the annual meeting of the National Council on Measurement in Education (NCME), Seattle.

Sheinker, A., Barton, K. E., & Lewis, D. M. (2004). *Guidelines for inclusive test administration 2005*. Monterey, CA: CTB/McGraw-Hill. Retrieved March 22, 2005, from http://www.ctb.com/media/articles/pdfs/general/guidelines_inclusive.pdf

Spearman, C. (1904). “General intelligence” objectively determined and measured. *American Journal of Psychology, 15,* 201-293.

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large-scale assessments* (Synthesis Rep. No. 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved May 1, 2003, from http://education.umn.edu/nceo/OnlinePubs/Synthesis44.html

Thompson, S., & Thurlow, M. (2002). *Universally designed assessments: Better tests for everyone!* (Policy Directions No. 14). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved April 29, 2003, from http://education.umn.edu/NCEO/OnlinePubs/Policy14.htm

Thurlow, M., House, A., Boys, C., Scott, D., & Ysseldyke, J. (2000). *State participation and accommodation policies for students with disabilities: 1999 update* (Synthesis Rep. No. 33). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved January 1, 2005, from http://education.umn.edu/NCEO/OnlinePubs/Synthesis33.html

Wiley, D. E., & Haertel, E. H. (1996). Extended assessment tasks: Purposes, definitions, scoring, and accuracy. In M. B. Kane & R. Mitchell (Eds.), *Implementing performance assessments: Promises, problems, and challenges*. Mahwah, NJ: Erlbaum.

Willingham, W. W., & Cole, N. S. (1997). *Gender and fair assessment*. Mahwah, NJ: Lawrence Erlbaum Associates.

Notes

[1] This report draws upon earlier ECD work of Eric Hansen, Robert Mislevy, and Linda Steinberg (Hansen, Mislevy, & Steinberg, 2003) as well as the work of Mislevy, Steinberg, and Russell Almond at ETS in the formulation of ECD (Mislevy, Steinberg, & Almond, 2002). In some instances, the report also reflects more recent work (Hansen et al., 2005; Hansen & Mislevy, 2005).

[2] As can be seen from this description, as well as from the foregoing quote from Kane (1992), when we refer to the validity argument, we are generally referring to the claims, evidence, warrants, etc., and their interrelationships, rather than merely to the claims.

[3] These models allow one to anticipate whether good measurement would likely result from attempting to assess a particular targeted proficiency given a person's profile (including disability or language status, if applicable) under a certain set of testing conditions.

[4] The Web-related requirements of the Section 508 standards bear considerable similarity to the W3C Web Content Accessibility Guidelines.

[5] Work by Hansen et al. (2005) further distinguishes between the targeted proficiency (that is the actual target of inference) and focal KSAs that are essential parts of the targeted proficiency but are not necessarily enumerated in reporting.

[6] What Mislevy et al. (2003) referred to as *additional KSAs* may be compared to terms used by others. Consider, for example, a cluster of terms that have been referred to as *ancillary or enabling skill requirements* (Haertel & Linn, 1996, p. 63) or *ancillary abilities* (Aguirre-Muñoz & Baker, 1997; Wiley & Haertel, 1996). According to Aguirre-Muñoz and Baker (1997), ancillary abilities refer to the set of skills or abilities "required for successful completion of a task that are not explicitly part of what is assessed" (p. 13). Haertel and Linn argued, "If some examinees are deficient in a test's *ancillary* abilities, then it is biased against them" (1996, p. 63, emphasis in original). This usage of the term *ancillary abilities* seems similar to what Elliott and Roach (2002, p. 12) have termed *access skills*. This usage also seems consistent with what Hansen and Mislevy (2005) called *ancillary requirements*, but which more recently have been called *nonfocal requirements* (Hansen, Mislevy, & Steinberg, 2008). (Following the publication of Hansen and Mislevy [2005], the use of the term *ancillary*, as in ancillary KSAs, was replaced with the term *nonfocal*.) A key point worth

emphasizing is that the distinction between nonfocal and focal KSAs is driven by the definition of the targeted proficiency. On the other hand, whether or not a KSA is a requirement is driven by whether the KSA is necessary to perform well in an operational assessment setting. Ambiguity as to whether a KSA is focal or nonfocal suggests the need for increased precision in the definition of the targeted proficiency and/or an improved task design framework to sort out confounded skills. Notwithstanding the value of comparisons with other terms, the definition of additional KSAs as outlined by Mislevy et al. (2003) is sufficient for the purposes of this report.

[7] The specific task features that drive demand for reading comprehension ability are described in greater detail in the design pattern shown later in this document.

[8] In this report we will give the most attention to the potential for reducing demands for focal KSAs (thereby tending toward unfair advantage for the test taker). However, the potential for increasing the demand for focal KSAs (thereby tending toward unfair disadvantage for the test taker) is also an issue.

[9] Note that attention needs to be given to features of the task situation that go beyond administration of the test items themselves. Attention needs to be given to activities that may occur well before the day of administration (e.g., test-preparation information provided in test bulletins) as well as activities that may occur after administration (e.g., dissemination of score reports and guidance on their use). Thus, interpretation and use of scores is impacted by task features that are invoked, before, during, and after test administration (Hansen et al., 2003).

[10] In practice the determination of appropriate accommodations involves a range of considerations, among them, feasibility.

[11] Having a separate column for individuals with low vision is not an essential part of this application of design patterns, but was thought to be useful for highlighting similarities and differences relative to test takers without any disability.

[12] For the sake of clarity in this explanation, we ignore the many variations that could exist within the category of low vision.

[13] High reading comprehension demand is driven by task features that this example points to but could be described in yet greater detail. For example, specific sets of features could be

described that would elicit possible observations, such as deduction of the main idea of a passage, identification of the referent of anaphora, drawing conclusion from propositions in the text.

[14] If we wished to make a greater number of gradations of RC ability, then we would probably want to have different levels of RC demand, which would make RC demand a variable feature rather than a characteristic feature.

[15] This might include, for example, the font size of the test content and of the answer sheet.

[16] Essentially, assuming that granting unlimited time to all test takers is not feasible, then one approach would be to seek to grant each low-vision test taker an amount of time that would make testing speeded to a degree that would be similar to that experienced by nondisabled test takers using default testing conditions. A deeper examination of the extended testing time is needed but is beyond the scope of this report.

[17] Those demands for focal KSAs are typically intended to be imposed by characteristic features.

[18] Nevertheless, as may be surmised from this report, in existing assessments, the process of addressing what constitutes an appropriate accommodation can provide an occasion for revisiting and refining of the distinctions between focal KSAs and additional KSAs.

[19] In a test design that takes into account individuals with disabilities, the set of variable features will tend to be larger than it would be in a design that did not address the requirement of test takers with disabilities, since the process of accommodating the needs and requirements of such test takers will, by their variability, tend to transform characteristic features into variable features. Consider, for example, the feature of font size of the printed test. Because the default manner of presenting reading comprehension content is in 12-point font, it is tempting to think that a 12-point font size should be a characteristic feature of the test design. However, since we are attempting to address the access needs of partially sighted individuals, font size is better designated as a variable feature that may assume many values, at the very least, the two values of (a) 12 point and (b) greater than 12 point.

[20] There is no absolute guarantee that there exists a set of variable features that will satisfy all these constraints for a given individual, though many times there is a workable set.

[21] In this example, the three methods for implementing the readaloud accommodation (human reader, synthesized speech, or prerecorded audio) are assumed to be equally effective.

[22] The unfair advantage is indicated by a "poor" level in the targeted proficiency compared to "good" performance. This may also be termed a false-positive outcome.

[23] Some authors opt for a broadly encompassing definition of validity. According to Willingham and Cole (1997): "Validity is the all-encompassing technical standard for judging the quality of the assessment process. Validity includes, for example, the accuracy with which a test measures what it purports to measure, how well it serves its intended function, other consequences of test use, and comparability of the assessment process for different examinees" (p. 228). The social/educational objective of increasing inclusion of individuals with disabilities in an assessment system might be encompassed by this broader definition of validity. Yet regardless of the particular definition of validity that one is using, it seems clear that in addition to the core validity issue (whether a test measures what it intends to measure) there are other issues that that may influence a decision about what constitutes an appropriate accommodation.

[24] This question presumes that there are nuances of intended interpretation that are different from or finer than those that are embodied in the definition of the targeted proficiency.