Research Report

# Evidence-Centered Assessment Design for Reasoning About Accommodations for Individuals With Disabilities in NAEP Reading and Mathematics

*Eric G. Hansen*

*Robert J. Mislevy*

*Linda S. Steinberg*

*July 2008*

*ETS RR-08-*38

# Evidence-Centered Assessment Design for Reasoning About Accommodations for Individuals With Disabilities in NAEP Reading and Math

Eric G. Hansen
ETS, Princeton, NJ

Robert J. Mislevy
University of Maryland, College Park

Linda S. Steinberg
Consultant

July 2008

As part of its nonprofit mission, ETS conducts and disseminates the results of research to advance quality and equity in education and assessment for the benefit of ETS' constituents and the field.

ETS Research Reports provide preliminary and limited dissemination of ETS research prior to publication. To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html



**Abstract**

Accommodations play a key role in enabling individuals with disabilities to participate in the National Assessment of Educational Progress (NAEP) and other large-scale assessments. However, it can be difficult to know how accommodations affect the validity of results, thus making it difficult to determine which accommodations should be allowed. This study describes recent extension of evidence-centered assessment design (ECD) for reasoning about the impact of accommodations and other accessibility features (e.g., universal design features) on the validity of assessment results, using examples from NAEP reading and mathematics. The study found that the ECD-based techniques were useful in analyzing the effects of accommodations and other accessibility features on validity. Such design capabilities may increase assessment designers' capacity to employ accessibility features without undermining validity.

Key words: Accommodations, validity, reading, mathematics, universal design, National Assessment of Educational Progress

**Acknowledgments**

This document is almost identical to an unpublished paper (Hansen & Steinberg, 2004) that was commissioned by the National Research Council Committee on Participation of English-Language Learners and Students with Disabilities in NAEP and Other Large-Scale Assessments and was based on 2003 presentation to the committee. The commissioned paper served as a resource to the committee in producing the book edited by Judith A. Koenig and Lyle F. Bachman titled *Keeping Score for All: The Effects of Inclusion and Accommodation Policies on Large-Scale Educational Assessments* (National Research Council [NRC], 2004, p. ix, chap. 6). The key changes from that paper to this one include the following: the name of Robert J. Mislevy was added to the list of authors, an abstract was added, the acknowledgements have been expanded; the term *nonfocal* replaces almost every instance of the term *ancillary* (in keeping with a change in terminology), an unnecessary reference has been deleted, and minor editorial corrections have been made. The authors acknowledge: Russell Almond, Linda Cook, Daniel Eignor, Deanna Morgan, and other reviewers of various versions of this document; Kim Fryer and other editors of the document. Remaining errors are those of the authors.

## Table of Contents

## List of Tables

## List of Figures

Page

## Introduction

There is wide agreement that the availability of appropriate accommodations is an important key to fostering high levels of inclusion of individuals with disabilities in large-scale assessments such as the National Assessment of Educational Progress (NAEP). Accommodations should overcome accessibility barriers without giving the student receiving the accommodation an unfair advantage. However, it is sometimes difficult to determine whether an accommodation is appropriate or not. For example, there are significant differences in assessment practices regarding the use of the *read-aloud* accommodation (i.e., having test content read aloud to the student by a live reader, synthesized speech, or prerecorded audio). Even though the read-aloud accommodation is not permitted in the NAEP reading assessment, it is permitted in the reading assessments of some state-administered achievement tests (U.S. Department of Education, 2003).[1, 2] (The read-aloud accommodation is allowed in some other NAEP assessments, including the mathematics assessments [National Assessment Governing Board, 2003].[3])

This discrepancy in practice regarding the read-aloud accommodation in assessments of reading is important. For example, without such an accommodation a student who is blind and does not read braille may be unable to participate in the NAEP reading assessment. In addition, individuals with severe dyslexia or some other print-related disability that prevents or otherwise greatly hinders their access to visually displayed text might also be excluded or have scores with compromised validity. The explanation for the NAEP practice regarding the read-aloud in reading, according to the *Reading Framework for the 2003 National Assessment of Educational Progress*, hereafter referred to as the 2003 NAEP reading framework, is that "because NAEP is a reading comprehension assessment, test administrators are not allowed to read the passages and questions aloud to students" (National Assessment Governing Board, 2002, p. 3).[4] This explanation is not a detailed rationale, but does suggest that allowing a read-aloud accommodation would invalidate results for individuals receiving it.

How does one begin to resolve such differences of procedure, keeping in mind the need for maximizing *inclusion* of special populations as well as the need for ensuring *validity* of inferences arising from accommodated assessment administrations? What kinds of knowledge are relevant and how can one relate the various pieces of knowledge to inform accommodation-related decisions? This report seeks to provide a basic, yet coherent, framework for evaluating

the validity of assessment accommodations that can then be used to reconcile the need for inclusion, while safeguarding validity. Among the contributions that this framework seeks to make is a way of defining more specifically what one intends to measure. Clarity on this point is important in reasoning about what constitutes an appropriate or valid accommodation.

### *Evidence-Centered Assessment Design*

One potentially important avenue of thinking and research regarding accommodations for individuals with disabilities would seek to lay foundations for an *evidence-based* approach, drawing upon insights from fields such as evidentiary reasoning and educational measurement. The essential idea is to lay out the chain of reasoning—the underlying rationale from data to claims—for practices that may be widespread and familiar, yet remain largely unexamined. These practices typically may be quite successful; however, because the reasons they work remain tacit, improvement in response to changing technological, social, and legal environments is hindered. Evidence-based approaches have proved useful in fields such as law, science (e.g., medicine, natural resource exploration), and intelligence analysis. 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 (Cronbach & Meehl, 1955; Kane, 1992; Messick, 1989, 1994; Spearman, 1904). A recent contribution to that tradition is Evidence-centered assessment design (ECD), which was formulated at Educational Testing Service by Robert Mislevy, Linda Steinberg, and Russell Almond (2003). ECD seeks to make explicit the evidentiary argument embodied in assessment systems, thereby clarifying assessment-design decisions.

### *Progress*

Recently Hansen, Mislevy, and Steinberg (2003) described initial efforts to model the validity arguments of assessment that could be used in analyzing accommodated administrations for students with disabilities. The argument for an assessment might be summarized as including (a) a claim about a person possessing at a given level a certain targeted proficiency; (b) the data (e.g., item or test 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 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 has been compromised (Messick, 1989). An example of an alternative explanation for low scores by an individual with a disability would be 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., blind). An example of an alternative explanation for high scores would be that the accommodation eliminates or significantly reduces demand upon the examinee for some aspect of the targeted proficiency.

This approach used by Hansen, Mislevy, and Steinberg (2003) makes use of Bayes nets that model the validity argument of assessments, including those parts of the argument that involve accommodations. A Bayes net consists of a set of variables, a graphical structure connecting the variables, and a set of conditional distributions. One adds *evidence* to a Bayes net by setting variables to particular values.[5] Adding evidence can take the form of either (a) observing the values of certain variables and wanting to study 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. Changes made to these values propagate according to Bayes Theorem, yielding posterior (post-setting) values for each of the other variables. By inputting characteristics of the person (knowledge, skills, and abilities [KSAs] intended for measurement as well as KSAs not intended for measurement) and characteristics of the assessment (e.g., the accommodation, the definition of the *targeted proficiency*, and specification of the factors influencing performance under specific operational settings), a user of the model can receive output in terms of the likely validity of the interpretations made on the basis of scores.[6]

One particularly noteworthy feature of this approach is the clear distinction that it draws between targeted proficiency and what is termed *effective proficiency*. Targeted proficiency is what one intends to measure and is defined by the KSAs that must be present in order for the individuals to be considered as having a *good* (or adequate or successful) level in the targeted proficiency. Targeted proficiency is similar, though more focused in meaning, to what is often referred to as the construct of an assessment.) On the other hand, effective proficiency, is essentially what one actually is measuring and is defined by the factors that actually affect performance in an operational assessment setting.[7, 8] Basically, the definition of the targeted proficiency is a matter of choice or intent, while effective proficiency is largely determined by

empirical observation or experience. By keeping these concepts distinct, it becomes possible to reason more rigorously about how well one measures what one intends to measure, a concept that lies at the heart of the idea of validity. To relate these notions to our earlier discussion, we would say that factors that result in deviations between an examinee's actual level in the targeted proficiency and their effective proficiency tend to be causes of invalidity and of credible alternative explanations for scores. The essential ideas here are not new; indeed, they lie at the heart of validity argumentation, dating back to at least Cronbach and Meehl (1955). Embretson (1983) foreshadowed their role in test development—building tasks more formally in accordance with the theory of what one intends to measure. The contribution of ECD is to provide a more complete framework and conceptual toolkit for sorting through the interconnected issues of purpose, test design, and inferential processes.

The initial efforts to build and apply the ECD framework in the disabilities context have focused mostly on reading comprehension (RC) and have been presented in a variety of forums in the last year. These include, for example, the annual meetings of the Association of Test Publishers (Hansen, Mislevy, Steinberg, & Forer, 2003) and the National Council on Measurement in Education (Hansen, Mislevy, & Steinberg, 2003). While this kind of modeling approach makes many simplifying assumptions, the approach appears to significantly increase the set of accessibility-related issues that can be addressed in a principled fashion, as opposed to relying on ad hoc, piecemeal solutions.

***Purpose***

The purpose of this report is to show how evidence-centered assessment design can be applied in the area of accommodations on NAEP reading and mathematics assessments for individuals with disabilities.[9] The report will walk through the creation and use of various models, in an almost instructional manner, beginning with very simple models and progressing to more complex ones. The reason to build more sophisticated and complex models is to be able to handle straightforwardly a greater range of assessment situations. Many of these situations involve accommodations and most involve one or more credible alternative explanations for good or *poor* performance. By understanding the interactions among the variables that make up a model, one can better understand and address a greater variety of threats to validity. For the sake of continuity, several of the models deal with the read-aloud accommodation, though some of the more sophisticated models can deal with other accommodations. A variety of examples are

offered as a demonstration or proof-of-concept that the models handle situations in a reasonable fashion.

Much of the focus will be on creating and using Bayes net models to evaluate the validity of inferences that arise from accommodated administrations of the assessments. The report will provide examples of the operation of the models, discussing the process of making inputs (e.g., accommodation, definition of the targeted proficiency, a set of person characteristics) and examining the outputs (likely validity of the interpretations). It should be emphasized that as in any application of Bayes nets, it is not the use of Bayes nets per se that determines the value of an effort but the insights into the substantive relationships that are embodied in the Bayes net. The variables and relationships represented in the proposed models and supporting processes integrate knowledge from several fields, including educational measurement, accessibility research, assistive technologies, and special education.

Although the focus is primarily on individuals with disabilities, the report also notes issues relevant for other special populations (e.g., English-language learners) as well as test takers in general.

The paper will include discussion of the limitations and challenges in using the approach as well as possible future use of the approach for NAEP or other large-scale assessments. This report draws heavily on the earlier work.[10]

***Intended Benefits***

This work examines the basic nature of the assessment argument for test takers with disabilities. It seeks to provide a common framework and language for incorporating these understandings from a variety of sources. It also seeks to develop reusable argument structures that would be applicable in a wide range of assessment design situations and does so using computer-based tools for facilitating assessment design thinking. While the ideas and tools are illustrated with fairly clear-cut examples, the approach is expected to be applicable to more complex and realistic assessment settings. What will be achieved here is the creation of a framework for thinking about assessment arguments that at the same time addresses accommodations issues, ties in with research on validity argumentation, and leads to practical assessment-design work.

This approach and this report have a number of limitations, many of which will be discussed at various points throughout the report. The consequence of these limitations is that the

models of these assessment arguments should not be used mechanistically make key decisions about test design and use. The following major categories of limitations are worth emphasizing up front.

The first limitation pertains to the approach itself. Specifically, this approach does not overcome the inherent impossibility of overcoming all error or invalidity in assessments. Indeed, assessment-design decisions and accommodations policies are inevitably exercises in optimization, as any one alternative reduces some sources of error but opens the door to others. The work of validation is never fully complete; therefore, at some point, it is up to human beings to decide a course of action—be it an accommodation-related decision or other assessment design or implementation decision. Because threats to validity can come from anywhere, it is impossible to specifically identify or model all of them. Furthermore, any index or indicator of validity is a limited summary of information about the assessment argument under a particular condition or set of conditions. Running the machinery of a Bayes net model may indicate whether the results of administering the assessment in that set of conditions is likely to yield valid or invalid results. Yet for reason such as those just cited, the degree of likelihood is not precisely known.

The second limitation pertains both to the approach and to this particular study. Specifically, the models used in this paper will make many simplifications and assumptions, which further contribute uncertainty. Some simplifications are made by choice and others by necessity. For example, this report will attend to some key features of reading and mathematics while mostly ignoring other features. We choose to work with more straightforward models in this presentation because they rely on knowledge that is well known or easily grasped and therefore can be better communicated. It is sometimes necessary to make simplifications because of practical limitations of knowledge, time, and resources. Acquiring research knowledge about disabilities and accommodations is often challenging because of the great diversity of disabilities, accommodations, and the relatively low incidence of certain disabilities. Model creation is a *knowledge-consuming* exercise. Simplifications, assumptions, and guesses need to be made where knowledge about the true state of reality is lacking for any reason Fortunately, the explicit nature of the modeling enterprise gives us an opportunity to describe and examine key assumptions and simplifications so that users of model-based results can make informed judgments about how to use the results of the models.

Nevertheless, even with simplifications and assumptions, we believe that this activity can increase the set of situations that can be addressed in a principled and effective way. Thus, despite these limitations we hope that this work can illuminate the nature of decisions and help assessment designers think through the sometimes-competing goals of assessment designs. It is also hoped that the approach will also help assessment planners think creatively about how to include individuals with a greater variety of disabilities by expanding the list of accommodations available for consideration. Furthermore, as referred to later in this report, it is hoped that this approach will help determine the nature and scope of the features that are made available under the heading of *universally designed assessments*.

### *Suggestions to the Reader*

This report is intended to convey its central message to the reader in the main body of the report. Material in the footnotes and appendices is supplemental and is made available for those individuals who are interested in additional details and nuances.

### *Overview of the Sections*

Following is an overview of the contents of the various sections.

*Basic concepts.* This section begins with a brief overview of the ECD framework and then provides basic background in argument structure and its applications to accommodation-related situations. The basic structure of an argument is illustrated through a Toulmin diagram (Toulmin, 1958).

The Basic Concepts section explains the concept of focal KSAs (KSAs that are essential constituents of the targeted proficiency) and nonfocal KSAs (those that are not). It also explains how KSAs may be either required or not required for good effective proficiency (i.e., in a specific set of operational assessment conditions). It notes how a KSA for a specific assessment situation may be categorized with respect to requirement status as (a) a *focal* requirement, (b) a *nonfocal* requirement, or (c) not a requirement. The importance of these terms is explained. For example, they allow one to characterize a basic goal of an accommodation (e.g., to reduce or eliminate nonfocal requirements so that deficits in nonfocal KSAs are not the cause of poor performance). The Basic Concepts section introduces the example of Sue and low vision, which figures prominently in the next two sections.

*Another way to represent the assessment argument***.** This section (a) discusses how Bayes nets can be used to model the validity argument of assessments, including those involving accommodations; (b) refines our understanding of KSAs, treating them not merely as present or absent, but sometimes as having multiple levels; (c) walks through the creation of a simple Bayes net model; (d) discusses several possible definitions of validity; and (e) explains some of the assumptions underlying the model. It should be noted that this report takes the approach that there may be no perfect index of validity. N evertheless, this section discusses three specific validity criteria that seem useful in the context of this modeling activity.

*A more complex example—Blind and read-aloud.* This section discusses the example of a test taker we will call Tim. Tim is blind and requests a read-aloud accommodation on an assessment of reading. This section examines the NAEP prohibition against the read-aloud accommodation in reading and describes the rationale provided by the NAEP reading framework document.

*A more complex example.* This section introduces a general schema for representing assessment accommodation situations, explains some of the strengths and limitations of that representational approach, and then shows show the argument structure can be more richly represented in a Bayes net. The section also discusses what we term *psycho-physical modeling*, which is a technique for building Bayes net models of any significant degree of complexity and utility in this domain.

*A simple model for NAEP reading and mathematics.* This section describes a single Bayes net model for NAEP reading and mathematics, based, in part, upon a review of the framework documents. This section explains the challenges in mapping directly from the frameworks to this application of ECD. The assumptions made and the course taken are described. This model consists of 34 nodes (variables) and handles five accommodations (large font, read-aloud, braille, and two kinds of dictionary). Six accommodation packages (each consisting of two accommodations) are available. The model allows one to specify hundreds of different examinee profiles, including those involving five specific disabilities—low vision, blind, deaf, dyslexic, neuropathic—or combinations of disabilities (e.g., deaf-blind-dyslexic-neuropathic). The model allows the user to specify two different definitions of the construct (reading, mathematics) and hundreds of different task performance settings (sets of testing conditions). Counting the variations in examinee profiles (including disability), task performance

situations (including accommodations), and definitions of the construct, the model allows one to analyze well over 100,000 situations in terms of the likely validity of the scores that would result.

This section describes briefly how this modeling activity relates to empirical (including quantitative) research. Finally, this section describes an example of how the modeling approach could be useful in a situation in which an assessment had greater-than-intended requirements for vocabulary knowledge for all examinees and how certain kinds of dictionaries may be either well-suited or ill-suited to addressing this issue. The dictionary example points to a wider set of issues than those that can be addressed through accommodations per se and points to other strategies, including universal design, for addressing them.

*Discussion and conclusions* The discussion portion of this section reviews some key examples and how they were modeled. This portion summarizes the approach and outlines the key steps. It also seeks to place this work with accommodations within the context of a large set of strategies for dealing with unmet nonfocal requirements. Among these strategies are increasing test taker capacity in nonfocal skills, universal design of assessments, and changes to the definition of the targeted proficiency. The section also asserts the relevance of this modeling approach in thinking through issues related to this wider set of strategies. The conclusion portion of this section provides some closing remarks and makes recommendations for possible next steps.

## Basic Concepts

### *The ECD Framework*

In seeking to apply ECD to the area of assessment accommodations, we will give considerable attention to a number of different forms of knowledge representation that highlight various issues and relationships in the design and analysis of assessments. We take the approach that there are ways of representing knowledge that can help put the issues of assessment accommodations into a more comprehensive validity framework and that can help us optimize the competing goals behind assessment accommodations. By developing a deeper and more shareable way of representing the assessment argument for test takers with disabilities we can better remove accessibility barriers faced by people with disabilities and at the same time safeguard the integrity and validity of our assessments.

A *knowledge representation* is a structure for expressing, communicating, and thinking about important entities and relationships in some domain (Markman, 1999).[11] Familiar examples include blueprints, flowcharts, and chess diagrams. Knowledge representations are surrogates for something else (e.g., a real-world situation or class of situations). Knowledge representations capture some entities and relationships while ignoring others. Knowledge representations are useful when they highlight important relationships and make them easier to work with:

- They facilitate analogical reasoning across problems and domains. For example, the knowledge representations in this report may help identify ways in which accommodations in reading, mathematics, and other subject areas are similar and dissimilar.
- They make it easier to acquire and structure information. Knowledge representations can help elicit knowledge from domain experts.
- They can facilitate working together. Objects of the argument structure can facilitate sharing, reusing, and repurposing elements and processes in assessment.
- They are significant in planning. What will a solution have to look like? What elements in assessments must hold and what elements can vary substantially?
- They encourage the use of overlapping knowledge representations to coordinate work in complex systems. One can develop multiple knowledge representations suited to different communication or other needs.

Knowledge representations come in many forms but may be referred to as models, objects, structures, layers, maps, and schema.

### *Layers of an ECD-Based Assessment Design*

ECD describes layers of assessment design and delivery in order to sort out the kinds of thinking and activity that take place at different points in an assessment system. The layer that represents the delivery of an operational assessment deals with items, scores, and assessment conditions—this is what examinees experience, and this is where accommodations occur.

As part of the delivery layer, there are three phases of assessment system activity. These phases are the pre-administration phase, the administration phase, and the post-administration

phase. The pre-administration phase and the post-administration phase of assessment system activity are important for issues related to accommodations. For example, during the pre-administration phase, one needs to arrange for appropriate assessment accommodations. During the post-administration phase of assessment system activity one needs to provide scores and guidance on their interpretation and use.[12]

The design layer—called the conceptual assessment framework (CAF)—is where the structures for the operational assessment such as psychometric models and test specifications, are laid out. Such models and specifications are a central concern of the assessment planner. The most prominent models of the CAF are the following:

1. The student model—the learner characteristics that one wishes to assess
2. The evidence model—procedures for task scoring and for updating beliefs about student-model variables
3. The task model—specifications for the task performance situation

The models of the CAF provide specifications for the administration phase of the assessment system activity, but, as noted earlier, when test takers with disabilities are involved, one also needs to pay particular attention to variables that are outside the CAF. These three models will be discussed further in this report.

In the modeling activity of this report, we will be working primarily at a higher, earlier, and less visible layer called the *domain model*, without which inferences based on scores cannot be adequately interpreted. The domain model layer is where the essential assessment argument is cast and, indeed, might be viewed essentially as the *argument* layer. At the domain model layer we are working with a wider set of variables than those we are concerned with in the CAF.

Key variables of the domain model essentially evolve into CAF variables; this is a matter of determining the exact grainsize and nature of variables in psychometric models for capturing the distinctions among examinees that are needed to support the purpose of the assessment. The details of the evolution are unimportant but the mapping between elements is important. For example, the domain model variable that we refer to as effective proficiency corresponds to the key variable of interest in the student model of the CAF.[13] *Task features* in the domain model map to *task model variables* in the *task model* of the CAF.[14]

A key purpose for discussing this mapping between the domain model and the CAF are to show that the domain model covers more territory than the CAF. One needs to think about that larger territory, especially with individuals with disabilities. For example, standard psychometric models deal simply with the relationship between the student model variable (effective proficiency) and the item or test scores, both of which are part of the CAF. This helps us realize that psychometric models, despite their central role in educational measurement, are not capable of enabling inferences based on student scores without this more encompassing domain model (argument) layer—particularly when individuals with disabilities are involved. Indeed, the attention we give to the domain model is in large part a consequence of the fact that one cannot take for granted that certain test takers with disabilities have certain skills that can be or are assumed for individuals without disabilities and for whom the ordinary conditions of administration have proven adequate or effective. Interpretations that we made based on test scores for individuals without disabilities require additional deliberate consideration of the domain model.

See Appendix A for more information about how this approach of this report fits into the ECD framework.

***Basic Argument Structure***

Let us consider the basic argument structure, which is closely adapted from Stephen Toulmin (1958), who was particularly interested in inductive reasoning, also often called inferential reasoning.[15] The following are key parts of the argument structure, as shown in Figure 1:

- Claim—a hypothesis or conjecture; the thing to be proved
- Data—information that becomes evidence when its relevance to the claim is established by a warrant
- Warrant—a generalization that tells why the observation should change our belief about the claim (The warrant permits the inference from data to claim. The warrant may be thought of as an elaboration of the arrow that leads from data to the claim in Figure 1.)
- Backing—information that supports the warrant, just as the data supports the claim

- Alternative explanations—propositions that weaken or rebut the argument.
- Rebuttal data—information that supports the alternative explanations.

To summarize, the core argument leads from the data to the claim. The warrant (with backing) permits (or licenses) inference, while alternative explanations (backed by rebuttal data) may tell about possible weaknesses in that inference.

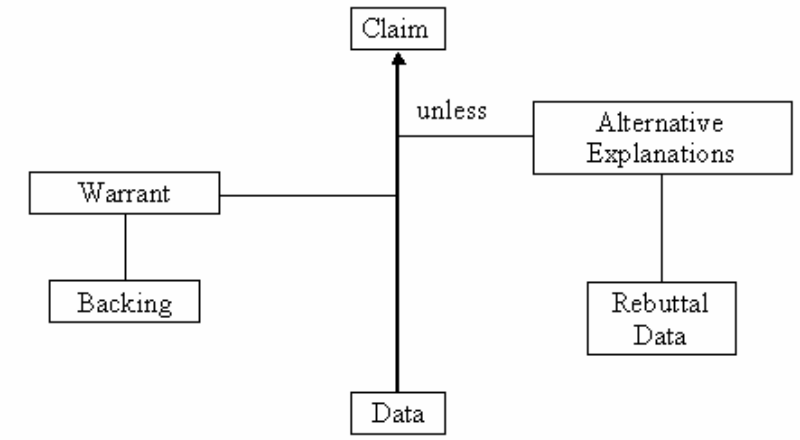


***Figure 1.*** **Basic argument structure.**

***Example 1: Sue and Low Vision***

Let us now consider an application of this argument structure in the context of an educational assessment situation. Let us suppose that we receive data that indicates that, in response to the presentation of a printed reading passage, a student we will call Sue incorrectly identified the main idea of a reading passage. This leads us to claim that Sue has poor RC. This claim is permitted by a warrant that says:

> IF a student has good RC,
> THEN she will probably answer a main idea item correctly;
> AND
> IF a student has poor RC,
> THEN she will probably answer a main idea item incorrectly.[16]

In this case, the warrant has backing in the form of accumulated theory and experience. One part of the backing might come from the knowledge of teachers who, for many years, have observed

the work and behavior of students who they know have good (and poor) RC proficiency and have performed correspondingly well (and poorly) on tasks such as those in the test. This assumes, for example, that by asking a student to indicate the main idea for a reading passage, one can elicit evidence about a student's reading comprehension ability. Another part of the backing might concern measurement theory and the applicability of specific techniques for specifying the statistical relationship between item scores and estimates of RC proficiency. Further backing could come from knowledge about different test takers in the target population and the conditions (presentation format, time allowed, type of item, etc.) under which items need to be presented in order to give students a fair opportunity to demonstrate RC skill.

Now let us suppose that we receive rebuttal data, additional information that appears to weaken the argument. Specifically, Sue appears to understand our e-mail exchanges, and we notice that she uses a large font size. This rebuttal data supports the idea—the alternative explanation—that Sue has low vision and was *not* able to access the test content because it was displayed in a font size that was too small for her to read, thereby causing unduly poor performance (we know that the test was administered using a regular-sized font). This situation is represented in the Toulmin diagram in Figure 2.

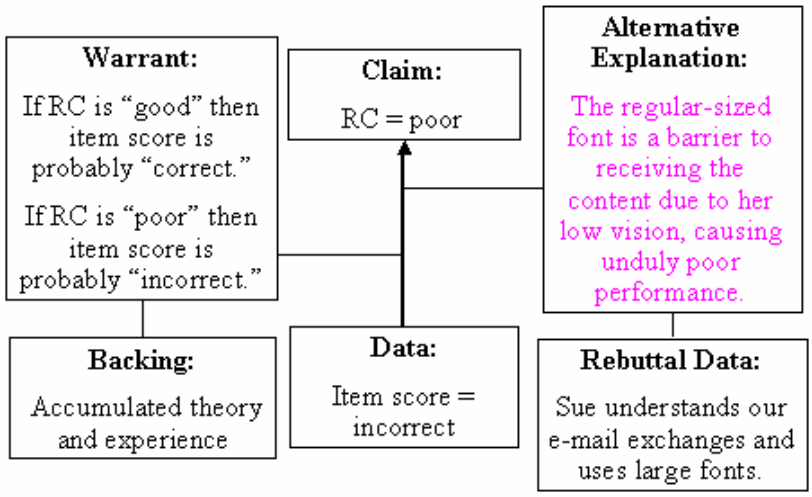


***Figure 2.*** **Toulmin diagram for the argument for Sue.**

Suppose we follow up with Sue and determine that she does indeed have low vision and can benefit from a large font size. We decide to test Sue again, this time using a large font size as an accommodation. In this case, Sue answers the item correctly and, therefore, our claim is now

that that she has good RC proficiency. We now have greater confidence in our claim than before since we have addressed the alternative explanation.

We see that providing appropriate assessment accommodations, such as presenting material in larger-than-usual font, is one way of addressing alternative explanations for poor performance. Sue originally answered the question incorrectly; while the initial explanation for that incorrect score was that she lacked skill in reading comprehension, there was a credible alternative explanation for that incorrect score. By changing the task performance situation for Sue (i.e., by granting an accommodation of large font), we could address this alternative explanation for poor performance. When we address alternative explanations, the assessment argument becomes stronger.

Ideally, of course, one should address likely alternative explanations as early as possible, preferably before they become problems. For example, by determining test takers needing accommodations in advance of test administration, one can avoid the difficulty encountered in the original assessment of Sue. Thus, we see that the basic argument structure is able to represent situations involving a test taker with a disability, including situations both with and without accommodations.

It should be noted, as shown in Table 1, that many alternative explanations for poor performance have nothing to do with disabilities. A test taker may have gotten a poor night's sleep the night before the test. Or the test taker may have not spent time with the familiarization materials before taking the test. A testing or assessment organization does not bear the full responsibility for addressing all these alternative explanations for poor performance, but it can take steps to address many of them. For example, by providing information in test bulletins that guides candidates in their long- and short-term preparation for testing, a testing organization can minimize the likelihood of credible (and true) alternative explanations.

What are the major kinds of alternative explanations that testing accommodations might address? This can be inferred from an examination of the commonly cited kinds of testing accommodations—presentation format, response format, timing, and setting as shown in Table 2 (Thompson, Thurlow, & Moore, 2002; American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME], 1999, p. 103).[17]

**Table 1**

***Examples of Alternative Explanations for Poor Performance***

| Alternative explanation | How a testing organization might address the issue |
|---|---|
| Test format does not permit reception of test content due to a disability | Before administering the test, identify an alternative format that will overcome the barrier to reception. |
| Test format is unfamiliar | Provide adequate text preparation materials. Require prior experience with a format when considering requests for an accommodation. |
| Student is sleep deprived | In the bulletin, encourage getting a good night's sleep before the test. |
| Student is emotionally upset by a family argument | If the issue is identified soon enough, allow the student to take the test at a different time. |

**Table 2**

***Four Kinds of Testing Accommodations and the Alternative Explanations for Poor Performance that Testing Accommodations Can Address***

| Kind of accommodation | Example | Alternative explanation (for poor performance) that the accommodation might address |
|---|---|---|
| Presentation format | Human reader, braille version, large-print version | Cannot receive the test content due to sensory disability |
| Response format | Scribe, mark answers in booklet | Cannot record answers due to physical disability |
| Timing | Extra testing time, frequent breaks | Lacks sufficient time |
| Setting | Special location, furniture, lighting, or acoustics | Cannot access the room<br>Cannot hear or see the proctor<br>Is distracted by others in room |

It is important to note that people generally request accommodations to address what they believe is an unfair disadvantage in taking the test under default (standard) conditions.

***Example 2: Spelling Disability and Spell Checker***

Let us now consider another example. Suppose we intend to measure the spelling skill of Carl, a student who has a spelling disability (dysgraphia). We receive a proposal that the student be allowed to use a spell checker on the spelling test. How should we respond to this request?

It seems clear, as shown in Figure 3, that the spell checker could benefit the performance of a person with a spelling disability; the spell checker would, thus overcome what is arguably an accessibility barrier. Yet, if the accommodation is approved, we can foresee a credible alternative explanation for good performance. Specifically, the spell checker would make it impossible to detect a student's poor spelling ability, thereby causing an unduly high (good) claim of spelling proficiency.

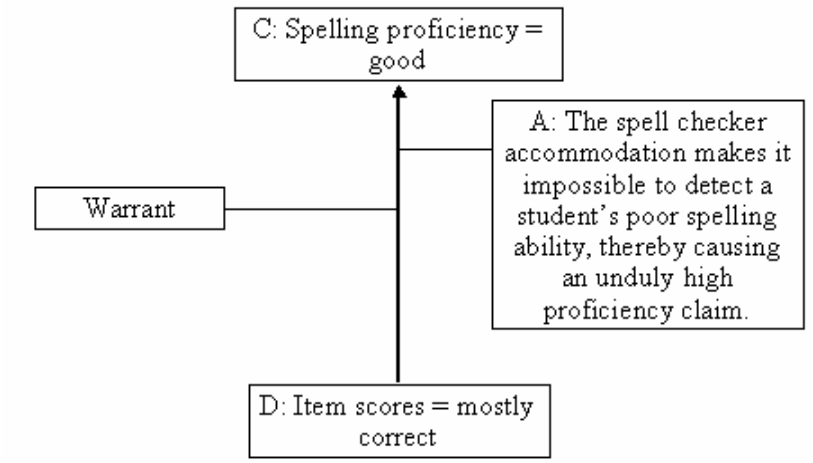


***Figure 3.*** **Spell checker accommodation: A credible alternative explanation for good performance.**

A testing organization might disallow such an accommodation, requiring the person to take the test under default conditions.[18]

As in the case of alternative explanations for *poor* performance, most of the alternative explanations for good performance pertain to test takers in general. An important example of alternative explanations for good performance involves forms of cheating, such as copying answers from someone else or receiving coaching in illegally obtained, test-specific content.[19] Table 3 shows some possible alternative explanations for good performance and how they might be addressed.

We see that testing accommodations are generally requested (and granted) in order to address unfair disadvantages, but that it is possible for an accommodation to cause an unfair advantage for the person receiving the accommodation.

Accommodations that are reasonable or appropriate should avoid both unfair advantages and unfair disadvantages. As Linn (2002) wrote: "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."[20]

**Table 3**

***Alternative Explanations for Good Performance***

| Alternative explanation | How a testing organization might address the issue |
|---|---|
| Testing accommodation eliminates or significantly reduces demand for some aspect of the targeted proficiency | Disallow the accommodation or find alternative forms of evidence |
| Test taker copied answers from a neighbor during the test | Tighten test center security |
| Test taker received coaching in illegally obtained test-specific content | Change item pool frequently<br>Test less frequently<br>Prosecute lawbreakers |

Alternative explanations, whether for good or poor performance, should be anticipated, if possible, and addressed before they become problems. We must keep in mind that alternative explanations (or threats to validity) can arise from virtually any quarter and that, as Messick (1989) has noted, validity evidence is never complete in a definitive way. However, we can learn to optimize our use of resources to address alternative explanations that are most likely or harmful.

How do we recognize which alternative explanations are most likely to need attention? Is it simply a matter of intuition and insight, or are there knowledge representations that can help lead us to recognize the issues most needing attention? A helpful step in developing improved representations is a common set of meanings and vocabulary.

### *Focal Versus Nonfocal Knowledge, Skills, and Attributes*

One of the most important distinctions that we can make concerns that between two different kinds of knowledge, skill, or other attribute of the student (KSAs)—that is, focal KSAs and nonfocal KSAs. These KSAs may be cognitive characteristics (*comprehend*, *know mathematics vocabulary*, etc.) or physical/sensory (*see*, *hear*, etc.). A focal KSA is a KSA that is defined by the assessment planner as an essential constituent of the targeted proficiency. More specifically, a focal KSA is one that an examinee must possess in order to be considered as having a good (or adequate or successful) level of the targeted proficiency.[21] Thus, the targeted proficiency is composed of one-or-more focal KSAs. A nonfocal KSA is any KSA that is not a

focal KSA. Note that it is the relation of a particular KSA to the definition of the targeted proficiency that makes it a focal KSA or a nonfocal KSA. That definition is fundamentally a matter of choice or intent—hopefully that choice is informed by theory and research, but it is a choice, nonetheless.

Let us consider an example of focal KSAs versus nonfocal KSAs for two hypothetical assessments—listening comprehension and reading comprehension, as shown in Table 4. For each of these assessments, each of nine KSAs (e.g., see, hear) is rated as being either a focal KSA or a nonfocal KSA. Obviously, the key challenge is to determine what are the focal KSAs; all other KSAs are, by definition, nonfocal KSAs.

**Table 4**

***Focal Versus Nonfocal Knowledge, Skills, and Abilities for Hypothetical Targeted Proficiencies***

| KSA | Assessment | |
|---|---|---|
| | Listening comprehension | Reading comprehension |
| See | Nonfocal KSA | Nonfocal KSA |
| Hear | Focal KSA | Nonfocal KSA |
| Decode | Nonfocal KSA | Focal KSA |
| Comprehend | Focal KSA | Focal KSA |
| Speak | Nonfocal KSA | Nonfocal KSA |
| Write using pencil or pen | Nonfocal KSA | Nonfocal KSA |
| Type on computer | Nonfocal KSA | Nonfocal KSA |
| Snow ski | Nonfocal KSA | Nonfocal KSA |
| Smell (olfactory sense) | Nonfocal KSA | Nonfocal KSA |

*Note.* KSA = knowledge, skills, and abilities.

In Table 4 we have posited, for example, that the ability to comprehend is a focal KSA (and therefore is an essential constituent) of both hypothetical targeted proficiencies. Focal KSAs other than comprehend are decode words from characters for reading comprehension and hear for listening comprehension. The KSAs of see, speak, write using pencil or pen, type on computer, snow ski, and smell (olfactory sense) are nonfocal KSAs for both reading comprehension and listening comprehension since they are not essential parts of either targeted

proficiency. One may argue that some of these KSAs—such as snow skiing and the ability to smell (the olfactory sense)—are so unrelated to an assessment of these academic proficiencies that they should not even be listed. The point of listing them here is to help emphasize the point that for any definition of a targeted proficiency, there are a potentially unlimited number and variety of nonfocal KSAs.[22]

A topic that is very important but will only be dealt with briefly concerns the process by which one generates definitions for targeted proficiencies. Among the relevant information for developing a definition would be information about the kinds of KSAs actually needed by individuals—including individuals with disabilities—who perform at a good/adequate/successful level in criterion situations. Such KSAs would be good candidates for focal KSAs. One would also want to look at the population for which the assessment is intended in order to confirm that there is appropriate variability in their ability in the focal KSAs. For example, it would not make sense to develop an assessment for a population of individuals whose focal KSAs were all at the highest level of all focal KSAs or all at the lowest level of all focal KSAs. However, regardless of the process gone through to define the targeted proficiencies, it is the definition itself that drives the distinction between focal KSAs and nonfocal KSAs.

### *Focal and Nonfocal Requirements*

The determination of whether a KSA is nonfocal or focal depends on the definition of the targeted proficiency. Yet, whether a KSA is required or not depends on whether the KSA is necessary for good performance in an operational assessment situation. KSAs may be either required or not required for good effective proficiency (good performance in an operational assessment situation). A KSA for a specific assessment situation may be categorized with respect to requirement status as (a) a focal requirement—a focal KSA that is required, (b) a nonfocal requirement—a nonfocal KSA that is required, or (c) not a requirement.

For example, we may have defined good reading comprehension as an individual having both good comprehension ability and good *decoding*. Thus, both comprehend and decode are focal KSAs. But it is a distinctly separate and equally important issue as to whether what we call reading comprehension items actually require any comprehension ability in order for the student to perform well. In other words, our intention is to assess whether the examinee possesses good reading comprehension ability (which we know requires good comprehension ability, in addition to good decoding ability); however, comprehension and/or decoding are focal requirements only

if the examinee must possess them in order to perform well in the actual assessment setting. This involves having task features capable of inducing those requirements as well as having appropriate ways of scoring the tasks (items) and the assessment as a whole.[23]

In order to know whether the focal requirements for a given assessment are appropriate, it is important, if not essential, to be able to have as a reference the focal requirements faced by individuals in a known case of valid measurement. Ordinarily, the most relevant reference point concerns the requirements faced by nondisabled individuals taking the assessment under default (standard) conditions. Translating a generally accepted principle of fairness and equity into these terms, we would say that the focal requirements of an assessment (whether or not it involves an accommodation) should be the same for all examinees, including nondisabled individuals taking the assessment under default (standard conditions). This implies that essentially the only thing we would ordinarily seek to modify through accommodation is the set of nonfocal requirements.

Consider Table 5, which examines possible focal and nonfocal requirements for our hypothetical assessment of reading comprehension. One can see from the third column of this table that, under default conditions, a student must use the KSAs of see, hear, decode words from characters, comprehend, and write using pen or pencil in order to perform well on the test of reading comprehension (i.e., to have good effective reading comprehension). This means that the task situation poses these requirements or demands to the examinee.

What kind of requirements are they? In order to know whether they are focal or nonfocal requirements, we must refer to the second column to see whether the KSAs were designated as focal or nonfocal based on the definition of the targeted proficiency. Of these five requirements, two are focal requirements (decode and comprehend) and the other three are nonfocal requirements. The fourth column indicates which these five requirements are focal requirements and which are nonfocal.

Does this set of requirements seem reasonable and desirable? First of all it seems reasonable that in order to obtain evidence about the focal KSAs, there needs to be a demand or requirement to exercise those focal KSAs. Furthermore, assuming that the assessment is administered in paper-and-pencil format, it is reasonable that in order to perform well, the student must be able to hear the proctor's instructions, see the paper test, and write with a pencil in order to record answers. It further makes sense that speaking, typing on a computer, using the sense of smell, and snow skiing are not requirements at all.

**Table 5**

***Requirements for Reading Comprehension Test Under Default Conditions***

| KSA | Focal versus nonfocal KSA (based on definition of the targeted proficiency) | Required to perform well (based on task demands in operational settings) | Kind of requirement | Task feature that generates the requirement under default conditions |
|---|---|---|---|---|
| See | Nonfocal KSA | Yes | Nonfocal requirement | Test is presented as visually displayed text |
| Hear | Nonfocal KSA | Yes | Nonfocal requirement | Proctor provides important information via speech alone |
| Decode | Focal KSA | Yes | Focal requirement | Frequency of words with multiple contiguous consonants |
| Comprehend | Focal KSA | Yes | Focal requirement | Complexity of sentence structures used in passage |
| Speak | Nonfocal KSA | No | n/a | |
| Write using pencil or pen | Nonfocal KSA | Yes | Nonfocal requirement | Students must write answers on answer sheet |
| Type on computer | Nonfocal KSA | No | n/a | |
| Snow ski | Nonfocal KSA | No | n/a | |
| Smell (olfactory sense) | Nonfocal KSA | No | n/a | |

*Note.* KSA = knowledge, skills, and abilities.

The table for an accommodated administration of the test would be structurally similar. Table 6 portrays the situations of a braille administration of the reading comprehension test. In this case, the focal requirements remain the same. The sense of sight is no longer a nonfocal requirement, but two additional nonfocal requirements have been added—reading braille (since the test content is presented via braille) and speaking (since the examinee must, let us suppose, dictate their answers to a scribe).

Even the simple tabular representation allows us to perform a simple analysis about the likely validity of the results obtained with this braille accommodation. We simply examine the person's profile of KSAs against the requirements column. Basically, if they can satisfy the

nonfocal requirements, then this supports the idea that they can be validly assessed. If they cannot, then they cannot be validly assessed. One advantage of this form of representation is its simplicity. However, when trying to capture some of the additional nuances arising in accommodation situations, this representational form has shortcomings and other representational forms become important.

**Table 6**

***Requirements for Reading Comprehension Test Using Braille Administration***

| KSA | Focal versus nonfocal KSA (based on definition of the targeted proficiency) | Required to perform well (based on task demands in operational settings) | Kind of requirement | Task feature that generates the requirement under accommodated conditions |
|---|---|---|---|---|
| See | Nonfocal KSA | No | n/a | |
| Hear | Nonfocal KSA | Yes | Nonfocal requirement | Proctor provides important information via speech alone |
| Decode | Focal KSA | Yes | Focal requirement | Frequency of words with multiple contiguous consonants |
| Comprehend | Focal KSA | Yes | Focal requirement | Complexity of sentence structures used in passage |
| Speak | Nonfocal KSA | Yes | Nonfocal requirement | Student must dictate answers to scribe |
| Write using pencil or pen | Nonfocal KSA | No | n/a | |
| Type on computer | Nonfocal KSA | No | n/a | |
| Snow ski | Nonfocal KSA | No | n/a | |
| Read braille | Nonfocal KSA | Yes | Nonfocal requirement | Test content is presented via braille |

*Note.* KSA = knowledge, skills, and abilities.

To summarize, thinking through testing accommodations in the context of an assessment argument necessitates an understanding of the definition of targeted proficiency so that we can distinguish between nonfocal and focal KSAs. We combine our knowledge about the focal/nonfocal distinction with our empirical knowledge about the levels of skills actually needed

to perform well in operational conditions in order to know what the focal and nonfocal requirements of the assessment situation are. Our empirical knowledge will also inform our plans for the task (e.g., item) features that will drive the focal and nonfocal requirements for various physical, sensory, and cognitive KSAs during actual performance situations. It is important to identify structures that help us organize the many pieces of relevant knowledge in order to help us make wise accommodation-related decisions.

It is important to note that focal and nonfocal requirements can be excessive, which works against validity. As for excessive focal requirements, consider a test of reading where knowledge of vocabulary is a focal KSA; notwithstanding it being a focal requirement, the difficulty of vocabulary could be higher than appropriate. Focal requirements can be too low, which also works against validity. For example, in our test of reading, the difficulty of the vocabulary can be lower than appropriate for a particular target population. Nonfocal requirements (which might also be called nonfocal, but necessary, KSAs), while virtually always present in an operational assessment, can arguably never be low enough. Indeed, it is the presence of nonfocal requirements that individuals with disabilities have difficulty satisfying that is generally the raison d'être for testing accommodations.[24]

It is also worth noting that even very heavy or high nonfocal requirements will not harm or thwart inferences if one knows that all examinees in the population can satisfy these nonfocal requirements.[25] In principle, nonfocal requirements only harm measurement when they are not satisfied. Eligibility rules for receiving accommodations and for taking an assessment at all are important considerations, because they provide one way for ensuring that all examinees can satisfy the nonfocal requirements associated with either the default administration conditions or some accommodated set of conditions. The efforts to use accommodations to increase inclusion of individuals with disabilities in NAEP and other large-scale assessments must focus on providing a broad range of accommodations that will allow each individual to receive the assessment in a manner in which the examinee's nonfocal abilities can satisfy the nonfocal requirements of the assessment so that their targeted proficiency, which is composed of focal KSAs, can be validly assessed.[26]

Thus, successful reasoning about accommodations requires a clear understanding of both the definition of the targeted proficiency (which allows the distinction between focal and nonfocal KSAs) and empirical knowledge about the demands or requirements driven by features

of the actual assessment situations.[27] Analysis of tasks can also help reveal the kinds of KSAs that are required for good performance (Sheehan & Ginther, 2001); such analysis can guide the empirical investigation of task demands. Yet neither the analysis of tasks nor the empirical investigation of task demands for certain skills can reveal whether any given requirement is focal or nonfocal, because it is the definition of the targeted proficiency that drives the distinction. An important role for a unified framework for reasoning about the validity of accommodations is to integrate knowledge from diverse sources (e.g., theoretical definitions of psychological constructs, cognitive analyses of task features, and empirical/quantitative studies of the factors affecting task performance).[28]

As noted earlier, virtually any assessment has nonfocal requirements. For example, any assessment that presents the examinee with a prompt or other set of instructions will typically involve a nonfocal requirement for sensory and other abilities to receive that prompt or instructions.[29] So the issue is not how to entirely eliminate nonfocal requirements, but rather to minimize or otherwise manage them to ensure that they are not the cause of an unfair disadvantage to the examinee. Which KSAs are involved and what levels are required are issues of test design; the same KSA can be a focal requirement in one assessment and a nonfocal requirement, or not a requirement at all, in another. Similarly, two assessments can be essentially identical except for their definition of the targeted proficiency; we will see an example of this in an analysis of the read-aloud accommodation for two different definitions of the reading comprehension targeted proficiency.

The distinction between the focal KSA and a nonfocal KSA is critical because our objectives during assessment design with respect to them are radically different. Essentially, our design objective relative to a focal KSA is to be able to ascertain or measure the targeted proficiency of which the focal KSA is a part. On the other hand, our key objective relative to nonfocal KSAs is to ensure that the examinee's levels in these KSAs are sufficient to satisfy the nonfocal requirements. Stated otherwise, one wants to ensure that deficits in nonfocal but necessary KSAs (i.e., nonfocal requirements) are not a cause of (i.e., an alternative explanation for) poor performance.

We can see that, while recognition of alternative explanations for good and poor performance might involve elements of chance or insight, an understanding of the distinction between focal and nonfocal KSAs can help anticipate alternative explanations and to determine

how to address them. For example, by distinguishing clearly between the focal and nonfocal KSAs for an assessment, one can then better think through the barriers that the nonfocal requirements of a test would pose for individuals with disabilities.

In the earliest stages of the ECD design process, such as during domain analysis, we begin with an act of imagination and seek insight. For example, we may imagine in what state an examinee must be, in order to perform optimally on an assessment. We think about them sitting at a desk, in a gym or classroom, or in front of a computer. We think of the media they need to receive and process information, the conditions in their environment, what they might need to know or be able to do technically, physically, socially, and intellectually. Thus, while we think of test takers as people, we start by also seeing them as possessing sets of skills or KSAs, perhaps not even differentiating between focal and nonfocal KSAs. Continuing the analysis, we think not only about individuals without any disabilities. Individuals who are blind, deaf, deaf and blind, dyslexic, have low vision, learning disabilities, cognitive disabilities, physical disabilities, and so on, may need to rely on a different set of nonfocal KSAs in order to receive the test content, to process it, and then to record their responses.

Proactive attention to test takers with disabilities during the early stages of assessment design pushes us to identify the nonfocal requirements that would exist for diverse test formats administered to individuals representing a range of disabilities. For example, for most academic tests administered under default (standard) testing conditions, the ability to see is a nonfocal requirement, since the examinee must be able to see the printed page to receive the test content and the sense of sight is not the target of measurement. The individual who is blind has a nonfocal KSA of sight that has a deficit, thereby requiring us to find another test format that will rely on another nonfocal skill that has no deficit. For example, by relying on the nonfocal KSAs of knowing braille codes and of being able to feel (sense of touch), the individual may be able to receive the test content.

An accommodation seeks to reduce or eliminate demand for one or more nonfocal requirements in which there is a deficit and instead rely on one or more KSAs in which there is no deficit. Selecting an appropriate accommodation involves matching the test taker's nonfocal KSAs to features of the task performance situation (task model variables), such that the nonfocal requirements of the task are those that can be satisfied by the individual with a disability.

By understanding the distinction between focal and nonfocal KSAs, one can make domain analysis more rigorous and lay a better foundation for domain modeling. Making basic distinctions between focal KSAs and nonfocal KSAs helps anticipate alternative explanations for poor performance (unfair disadvantages) and also provides alternative explanations for good performance (unfair advantages). This leads us to the following two rules of thumb:

1. Alternative explanations for poor performance can arise from (a) excessive nonfocal requirements (i.e., requirements for nonfocal KSAs at levels that exceed those possessed by the student) and (b) excessive focal requirements (i.e., requirements for focal KSAs at levels that exceed those experienced by students in known cases of valid measurement).

2. Alternative explanations for *good* performance can arise from insufficient focal requirements (i.e., requirements for focal KSAs at levels that are less than those experienced by students in known cases of valid measurement).

***A Return to Our Earlier Examples***

Let us review our first two examples—Sue and low vision, and Carl and spelling disability—but this time leveraging our understanding of the important distinctions between focal and nonfocal KSAs and requirements, while also thinking of these two rules of thumb.

Figure 4 pertains to the original situation with Sue and low vision, which involved no accommodation. Based on observation of an incorrect response to a test item, we reason through the warrant to claim that the targeted proficiency, RC, is poor. (Recall that the targeted proficiency is composed of one-or-more focal KSAs.) Yet an alternative explanation for the observed poor performance is that low vision does not satisfy the nonfocal requirement for sight caused by use of regular-sized font, thus resulting in unduly poor performance. This alternative explanation is arguably more precise and theoretically grounded than our previous alternative explanation, which was that the regular-sized font is a barrier to receiving the content due to Sue's low vision, causing unduly low performance. Providing the accommodation of large font size allowed that alternative explanation to be addressed.

Figure 5 pertains to Carl and spelling disability. We have a situation in which item scores are mostly correct, thereby supporting a claim that the targeted proficiency, spelling, is good. Yet if this occurs in a situation in which the spell checker accommodation is allowed, then we have

the credible alternative explanation that the spell checker reduces the focal requirement for spelling proficiency, thereby causing an unduly high proficiency claim. This alternative explanation is arguably more precise and theoretically grounded than our earlier explanation that the spell checker accommodation makes it impossible to detect a student's poor spelling ability, thereby causing an unduly high proficiency claim.

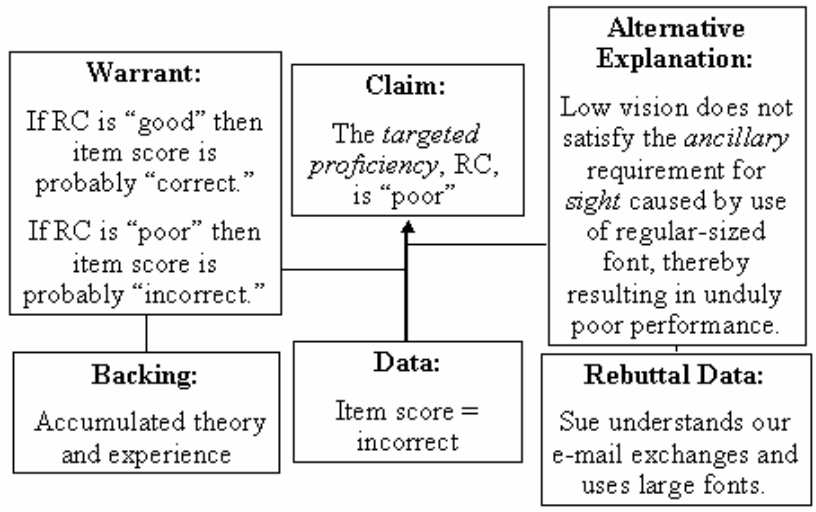


***Figure 4.*** **Sue and low vision: An alternative explanation for poor performance, referring to targeted proficiency and nonfocal requirement.**

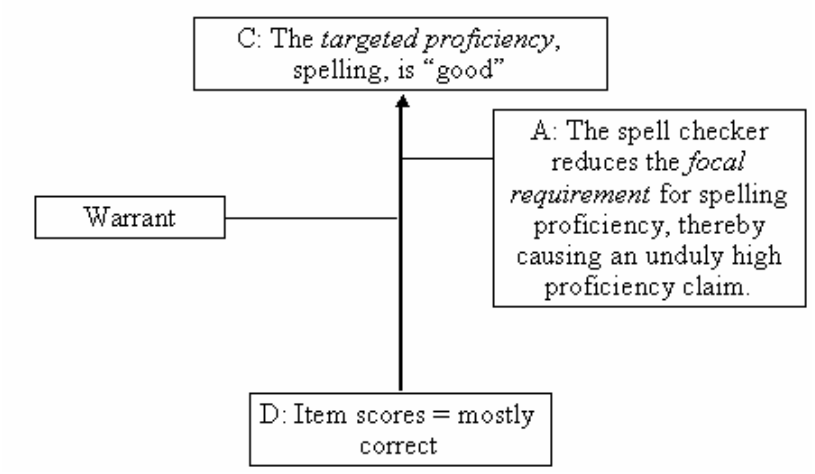


***Figure 5.*** **Spelling disability: An alternative explanation for good performance, referring to targeted proficiency (consisting of the focal knowledge, skills, and abilities [KSA] of spelling) and focal requirement.**

The distinction between focal and nonfocal KSAs is critical in our reasoning about accommodations. The use of the distinction between focal and nonfocal KSAs and requirements in the context of Toulmin diagrams may be considered a transitional knowledge representation—halfway between (a) the simple intuitive representation in our earlier Toulmin diagrams and (b) the more complex representations in Bayes nets that will be introduced.

While the Toulmin diagram is easily understood and is a valuable communication tool, other ways of representing assessment arguments can be useful in representing complex interactions between the argument variables and in pointing toward optimal solutions to situations involving competing goals. Such tools can help us anticipate problems early and address them before they occur.[30]

## Another Way of Representing the Assessment Argument

We will now consider another way to represent the assessment argument—Bayes nets, which are sometimes called belief networks. Bayes nets are useful in a wide range of situations in which we are interested in the interrelationships between many variables. We want to know how a given value of some variable affects our belief about the states of other variables that are either not yet observed or that are inherently unobservable. If it is possible to express the relationship among variables in terms of a joint probability distribution, then the machinery of probability gives rules about updating our belief when we learn about some other variables in the system and the implications for the remaining variables in the system. Bayes nets are a compact way of representing and updating these probability distributions.

One of the significant features of these Bayes nets is their flexibility. They can reason forwardly (deductively, from parent nodes to child nodes) or backwardly (inductively, from child nodes to parents nodes) with equal facility (Jensen, 1996).In this paper, we will use the tool Microsoft Bayesian Network Editor and Toolkit (MSBNx) for creating and displaying Bayes nets. Bayes nets give us a convenient way to represent the assessment argument in terms of a runnable model or models, which we can then share, analyze, and evaluate.[31]

A Bayes net consists of a set of variables, a graphical structure connecting the variables, and a set of conditional distributions. In MSBNx each node represents a random variable that can take any of two or more defined values; the node is displayed as an oval. The arc (arrow) points from the parent node to the child node (i.e., from the cause to the effect). The arrows may be seen as representing dependency, where the state of the child node depends on states of the parent

node(s). When constructing a Bayes net, one enters conditional probabilities, based on theory and experience.[32] Then, having constructed the Bayes net, one can set (or instantiate) variables consistent with some state of affairs of interest. This clamping action causes changes that propagate throughout the network in a manner governed by Bayes Theorem and yields posterior (post-clamping or after-clamping) probabilities for all the other variables.

A reason for using Bayes nets in this report is that it is a representational form that can express at once both the substantive assessment arguments that we have been discussing in terms of Toulmin diagrams and the design elements of the CAF, such as measurement models and task feature decisions. Bayes nets can clarify how the nuts and bolts of an operational assessment relate to the larger perspective of the assessment argument—relationships that are often tacit in ongoing assessments, but which are central to thinking through accommodations issues.

### *A More Refined Way to Define the Targeted Proficiency*

Before launching into the discussion of Bayes nets, let us consider a more precise way to define the targeted proficiency. Previously, we simply categorized each KSA as either a focal KSA or a nonfocal KSA, depending on whether it was a significant constituent in the definition of the targeted proficiency. That level of analysis seems to presume that each is an all-or-nothing or on-or-off KSA. But it is often helpful to distinguish between several levels of the focal KSAs. Among other advantages, defining several levels of KSAs allows us to define targeted proficiencies more precisely, a practice that is cited as being important to ensuring access to assessments by people with disabilities (Thompson, Johnstone, & Thurlow, 2002).[33]

Let us consider an example of how considering levels can lead to more precise definitions of the targeted proficiencies. Let us designate three levels of decoding ability—good, okay, and poor.[34] Suppose that we know that poor decoding is characteristic of the disability of dyslexia. (This is a simplified view of the disability but serves to illustrate the concept.) Let us now consider three possible definitions of reading comprehension, as shown in Table 7.

All three definitions of reading comprehension require that the individual have good (as opposed to poor) comprehension ability in order to be considered as having good (as opposed to poor) reading comprehension ability. These three definitions of reading comprehensions differ only with respect to the role of decoding. Under Definition B, the person needs a poor (or better)

level of decoding ability. Under Definition A, the person needs an okay (or better) level of decoding ability. Finally, under Definition C, the person needs a good level of decoding.[35]

**Table 7**

***Three Definitions of Reading Comprehension (i.e., Minimum Levels Required for Good Reading Comprehension)***

| KSA | Definition of reading comprehension | | |
|---|---|---|---|
| | B | A | C |
| 1. Comprehend (good, poor) | Good | Good | Good |
| 2. Decode (good, okay, poor) | Poor | Okay | Good |
| Decoding | Nonfocal KSA | Focal KSA | Focal KSA |

*Note.* The columns for the definitions are ordered in increasing level of decoding. This happens to be in a nonalphabetical order. Some of the same definitions (A and B) are used later in this report. KSA = knowledge, skills, and abilities.

Note that the last row in the table, which is not actually part of the definition, indicates whether decoding is a focal or nonfocal KSA in each of these definitions. Notice that decoding is a focal KSA for both Definitions A and C and is a nonfocal KSA for Definition B. The focal versus nonfocal designation is derived from (a) knowledge of the minimum level of decoding (one row up), (b) knowledge of the definitions of nonfocal and focal KSAs, and (c) knowledge of a convention that a KSA at its lowest level (in the case of decoding, poor) counts as being a nonfocal KSA. Specifically, in this document, a nonfocal KSA is one for which the lowest level of the KSA is sufficient for a person to be considered as having a good level in the targeted proficiency. The very lowest level of decoding (poor) is required for Definition B, meaning that decoding is a nonfocal KSA.

While we regard Table 7 as a reasonable way to define the targeted proficiency, some people find it confusing to have the designation poor level to be part of the definition. In order to avoid confusion, we will, where feasible, opt for the arrangement shown in Table 8, in which under Definition B, the designation poor is changed to n/a (for not applicable). Yet the information in Tables 7 and 8 is functionally equivalent in the context of our convention.

**Table 8**

***Three Definitions of Reading Comprehension, Showing n/a Instead of Poor for the Minimum Levels Required for Good Reading Comprehension***

| KSA | Definition of reading comprehension | | |
|---|---|---|---|
| | B | A | C |
| 1. Comprehend (good, poor) | Good | Good | Good |
| 2. Decode (good, okay, poor) | n/a | Okay | Good |
| Decoding: | Nonfocal KSA | Focal KSA | Focal KSA |

*Note.* KSA = knowledge, skills, and abilities.

A key point to note is that for some basic purposes related to accommodations, it is often enough to say simply that a KSA is focal or nonfocal requirement. Yet for other purposes, it is extremely useful to define the targeted proficiency in terms of the specific levels of focal KSAs that are necessary to be considered as having a good (or adequate or successful) level of the targeted proficiency. Note that Definition A and C of reading comprehension both have decoding as a focal KSA yet Definition C gives decoding a place of higher prominence in the targeted proficiency. Note that our use of only two or three levels for a KSA is a simplification. In many assessments, interest lies in an examinee's location along a continuum of proficiency.

### *Kinds of Knowledge To Be Represented in Bayes Nets*

One of the chief benefits of using Bayes nets to model the validity argument of assessments is their capacity to represent different parts of the argument and specify very precisely how those parts of the argument should interact. Our approach to modeling the argument keeps the various parts of the argument distinct from each other. For example, it keeps the definition of the targeted proficiency (construct) distinct from effective proficiency (which is determined by the actual demands and requirements on examinees during the operational assessment) and also distinct from the examinee's levels in various focal and nonfocal KSAs. One of the challenges in model building is getting these representations to interact in sensible ways.

Following are other kinds of knowledge that we need to represent in Bayes nets:

1. Knowledge about the requirements for nonfocal KSAs is affected by different presentation formats.

We may know that receiving test content via braille produces demands for several skills—a sense of touch, knowledge of braille codes, as well as decoding ability. On the other hand, receiving content via read-aloud requires a sense of hearing but essentially no decoding ability.

2. Knowledge about the amount of test-taker skills that are necessary to meet demands generated by the task situation.

For example, we may know that in order to meet a high demand for decoding in a reading test we would need a good (as opposed to okay or poor) decoding ability; to satisfy a medium demand we would need an okay (or better) level of ability; or, to satisfy a low demand, we would need only a poor (or better) level of ability. This kind of mapping between KSA levels is integral and routine to the simple models used in this document.[36] Our convention is shown in Table 9.

**Table 9**

***Minimum Levels of Knowledge, Skills, and Abilities Needed to Satisfy Levels of Demand***

| Level of KSA | Satisfies this (or lower) demand. |
|---|---|
| Good | High |
| Okay | Medium |
| Poor | Low |

*Note.* KSA = knowledge, skills, and abilities.

Of course, this convention is applicable only where the KSAs and demands (or requirements) are both expressed in levels. Continuous student model variables might be useful; such an extension increases the complexity of the model but probably does not add sufficient insight to the handling of the accommodations issues to justify its use in this report.

3. Knowledge about how failure to meet a focal or nonfocal requirement affects performance in operational testing conditions.

For example, we may know that good eyesight is required to do well, that a lack of good eyesight will virtually always result in a poor effective proficiency. (Indeed, our notion of nonfocal requirements is based on the idea that failure to satisfy them results in depressed performance.)

***Creating the Model***

The next section begins walking through the development of a Bayes net. Figure 6 shows a very simple Bayes net created in the Bayes net tool. There is an arrow leading from the RC node to the score-on-item-1 node. RC represents the targeted proficiency. This graphic is intended to illustrate a concept that has been integrally associated with educational measurement: the idea that a latent (hidden) characteristic of a person—such as proficiency in reading comprehension—takes the role of a cause (parent) or driver of observable events, particularly scores on assessment tasks or items.

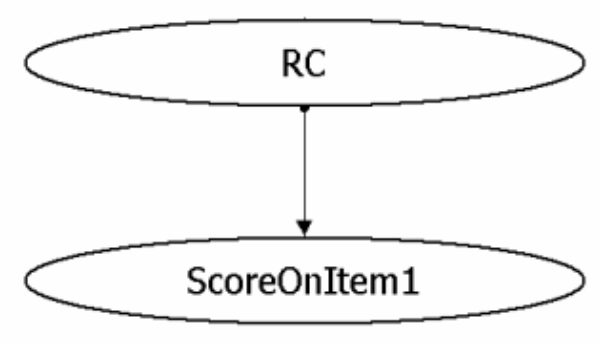


***Figure 6.*** **Reading comprehension proficiency causes the item score.**

Note that, like the Toulmin diagram, this Bayes net graphic includes both the claim (RC) and the data (item score); but, the arrow is pointed in the opposite direction from the arrow in the Toulmin diagram. Specifically, while the Toulmin diagram has the arrow pointing from the data (effect, item score) to the claim (cause, RC), this Bayes net has the arrow pointing from claim (RC) to the data (item score).[37] The arrows in a Bayes net signify a conditional probability distribution, of a child variable given all the variables that are its direct parents. These distributions expressed in the conditional probabilities in most of the examples in this report are generally logical (reflecting known relationships), though some are probabilistic (involving uncertainty).

We now add two additional nodes—one called effective RC and another called meet reception demand, as shown in Figure 7. This allows us to begin to represent an important source of alternative explanations.

The distinction between RC and effective RC may seem somewhat subtle, but it is very important. RC and effective RC are specific instances of the more general terms of targeted proficiency and effective proficiency, respectively.

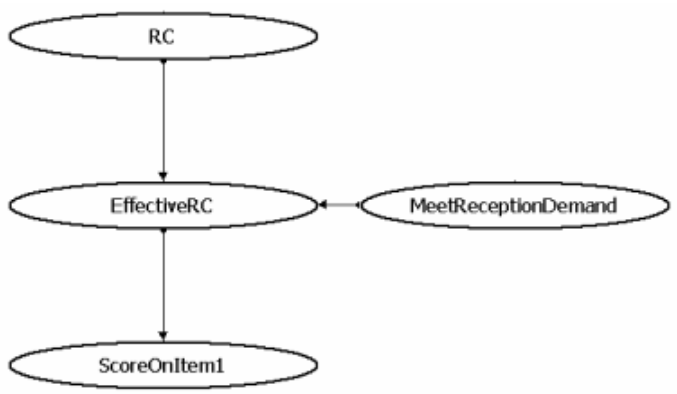


***Figure 7.*** **Four nodes.**

- *Targeted Proficiency—RC.* RC represents the targeted (but invisible) proficiency. It might be thought of as a perfectly measured criterion that reflects an intent of measurement, usually that of the test designer or assessment planner. It is desirable that the definition of the targeted proficiency be informed by a thorough understanding of the knowledge, skills, and other attributes (KSAs) that are actually required to perform well in criterion situations as well as the uses to which the scores will be put.[38] To keep matters simple, let us suppose that RC has two possible values (good and poor).
- *Effective Proficiency—Effective RC.* Effective RC is the targeted proficiency as operationalized in a given performance situation. Let us suppose that effective RC likewise has two possible values (good and poor).

Informally, we might say, in the context of this report, that RC is ability whereas effective RC is test performance. Effective RC is one of many possible operationalizations of RC. When the same operationalization serves well for all examinees, this distinction between RC and effective RC tends to fade into the background. Yet, when the choice of operationalization affects different examinees in different ways, the design decisions for measuring RC via effective RC become more crucial. As we shall see, it is at this junction that the assessment

planner must examine closely exactly what RC is meant to be, in order to serve the intended purposes of an assessment with data gathered under different operationalizations for different examinees.

Figure 7 shows effective RC having two parents, indicating that effective RC depends not only on their underlying (latent) proficiency (RC), but also on whether reception demand is met. The meet reception demand variable indicates whether the person is able to receive the test content. Let us suppose that meet reception demand has two possible values—yes and no. For example, based on experience, we know that if a person is completely blind and the test content is presented visually, then the value of meet reception demand is no.[39]

Before examining the conditional probabilities that we programmed into the Bayes net, let us examine briefly its behavior. As shown in Figure 8, when we set RC to good, and meet reception demand to yes, then effective RC is good.[40] One can see that this is reasonable, since if a person has good underlying reading comprehension ability and they can receive the content, their performance will probably be good.[41, 42]

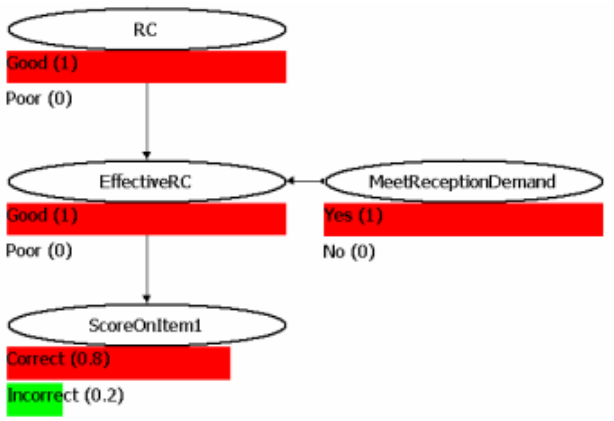


***Figure 8.*** **Good reading comprehension and meeting reception demand yield good effective reading comprehension.**

On the other hand, as shown in Figure 9, if we set meet reception demand to no and keep RC set to good as before, effective RC becomes poor instead of good. This result seems sensible, since even if someone's underlying RC ability is good, if they cannot receive the content—such as when a person who is totally blind is presented with a test using visually displayed text—then their test performance (effective RC) is likely to be poor.[43]

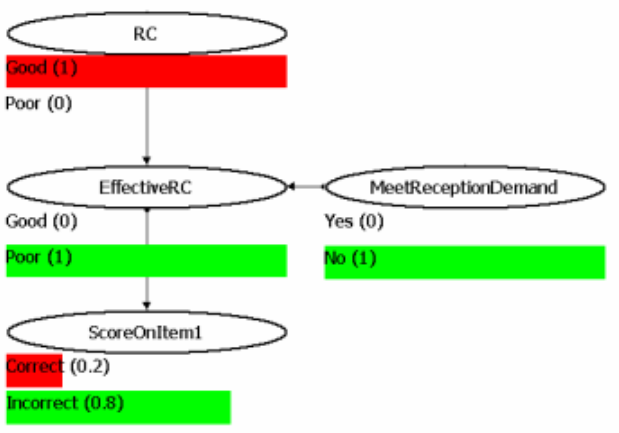


***Figure 9.*** **Good reading comprehension and *not* meeting reception demand yield poor effective reading comprehension.**

### *A Possible Definition of Validity*

It is important to note that comparing RC with effective RC yields one index or definition of validity. Arguably the central notion of validity is whether one is measuring what one intends to measure. Comparing RC with effective RC provides one way to operationalize this notion. If RC and effective RC both have the same value then we have validity (at least within the limits of that index). Specifically

- If RC = good and effective RC = good then we have a true-positive outcome and, therefore, have validity. This is the situation in Figure 8.
- If RC = poor and effective RC = poor then we have a true-negative outcome and, therefore, have validity.
- If RC = good and effective RC = poor then we have a false-negative outcome and, therefore, lack of validity. This is the situation in Figure 9.
- If RC = poor and effective RC = good then we have a false-positive outcome and, therefore, lack of validity.[44]

To the extent that the term *construct change* is used to describe situations in which there is a lack of validity—that is, a change when comparing what is intended to be measured with what is actually measured—then the *false* outcomes (false-negative and false-positive) can represent forms of construct change.[45]

While this index of validity is sufficient for some purposes, it has some notable limitations. For example, in the example we are studying, it is possible to have *true* outcomes, in this case, true-negative outcomes even when reception demand is not met. For example, suppose that an individual has poor underlying RC ability and their reception demand is not met because they are blind and the test content was presented to them as visually displayed text. In other words, they did not receive the test in an accessible format, although the performance result (effective RC as poor) would have been the same even if the test had been presented in an accessible format. In this case, one could argue that the interpretations arising from scores would be both invalid and unfair because the test content was inaccessible. Indeed, even if the score might be the same as would be encountered under proper conditions, the causal connection between the targeted proficiency (cause) and effective proficiency (and, hence, to scores [effect]) would have been damaged. The damage to movement through the causal chain is also reflected in damage to inferential reasoning flowing in the opposite direction (from scores to targeted proficiency), meaning that the damage adversely influences the inferences about the targeted proficiency that we can draw from item scores.

### *A More Stringent Definition of Validity*

A more stringent definition of validity that addresses this issue might define good or valid measurement as involving both (a) a true outcome (true-positive or true-negative) involving RC and effective RC and (b) reception demand being met. Under this definition, valid measurement for a person with a disability would require that reception demand be met in addition to having a true outcome. Not only must the outcome be correct, but the person needs to be able to receive the content (i.e., reception demand must be met) so that the performance could have gone either way, depending only on the examinee's level in the targeted proficiency.[46]

Having now seen the basic arrangement of the four nodes in this Bayes net and seen some of its behavior, let's take a look at the conditional probabilities that we have entered into the Bayes net to guide its behavior. When creating the Bayes net model, we coded into each node a set of prior probabilities. In a node that has parents, these prior probabilities are conditional probabilities for each of the node's values, given each possible combination of the values of its parents.

Consider, for example, Table 10. This display shows the conditional probabilities that have been entered for the effective RC node. On the left-hand side of the table we see all four

combinations of states of the two parent variables (meet reception demand and RC), and on the right-hand side we see probabilities associated with each of the two states of effective RC (good and poor). So, for example, in the first data row, we see that if meet reception demand = yes and RC = good, then there is a probability of 1.0 (i.e., 100%) that effective RC = good and, therefore, 0.0 (i.e., 0%) probability that effective RC = poor. From the other three data rows, one can see that where RC = poor or meet reception demand = no, there is a probability of 1.0 that effective RC = poor. Thus, the probabilities that we enter into the Bayes net capture the idea that the only way for someone to perform well in an operational setting (effective RC) is if they both possess the proficiency (RC = good) and are able to receive the content (meet reception demand = yes). [47]

**Table 10**

***Conditional Probabilities for the Effective Reading Comprehension Node***

| Parent node(s) | | Effective RC | |
|---|---|---|---|
| RC | Meet reception demand | Good | Poor |
| Good | Yes | 1.0 | 0.0 |
| | No | 0.0 | 1.0 |
| Poor | Yes | 0.0 | 1.0 |
| | No | 0.0 | 1.0 |

*Note.* RC = reading comprehension.

### *The Relationship Between Effective Reading Comprehension and the Item Score*

To be thorough, let us consider the relationship between effective RC and its child, the item score. As shown in Table 11, we have entered probabilities into the Bayes net to indicate that if a person's effective RC is good, then there is a 0.8 (80%) probability that their score on that item will be correct. [48] However, if their effective RC is poor, then there is only a 0.2 (20%) probability of their having an item score of correct.

What kind of theory or knowledge would give rise to these numbers? Let us suppose that this is a multiple-choice item with five possible responses and, therefore, a person could answer correctly about 20% of the time just by chance; that is captured by the 20% chance of having a score of correct, even with poor effective RC. Even under the best performance situations, we also know that a person can make careless mistakes or that other random errors may enter into a testing situation (such as a disruption at the test center happens when a fire engine rushes by with its siren blaring and distracting the test takers, etc.), thus accounting for a less-than-100%

probability of obtaining a score of correct, even when effective RC is good. The very simple error model we have employed in this table represents the larger issue of measurement error, which is a step of uncertainty between effective RC, however defined, and an observation meant to provide information about it.

**Table 11**

***Probabilities Correct or Incorrect Score, Based on Values of Effective Reading Comprehension***

| Parent node(s) | Score on Item 1 | |
|---|---|---|
| Effective RC | Correct | Incorrect |
| Good | 0.8 | 0.2 |
| Poor | 0.2 | 0.8 |

*Note.* RC = reading comprehension.

### *Adding Two Nodes: See and Font Size*

Now let us add two new nodes to the model—see and font size as parents of meet reception demand, as shown in Figure 10.

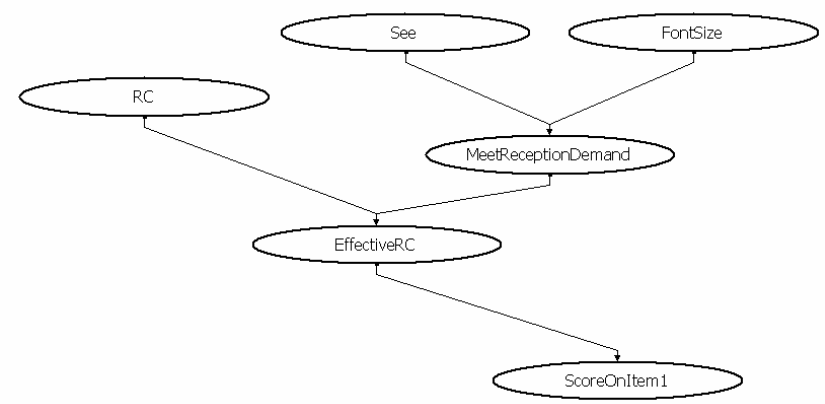


***Figure 10.*** **See and font size as parents of meet reception demand.**

We have seen that in order for a person with good reading comprehension (RC = good) to actually perform well (effective RC = good), reception demand must be met (meet reception demand = yes). When we entered conditional probabilities into the meet reception demand node, we specified the conditions under which that can occur. As shown in Table 12, a person with low vision (see = partial [low vision]) cannot receive test content with a regular-sized font. That is,

when font size is regular, the probability of reception is zero. On the other hand, that same person can receive test content where the font size is large (i.e., the probability of reception demand being met is 1.0).[49]

**Table 12**

***Conditional Probabilities for the Meet Reception Demand Node***

| Parent node(s) | | Meet reception demand | |
|---|---|---|---|
| See | Font size | Yes | No |
| Yes | Regular | 1.0 | 0.0 |
| | Large | 1.0 | 0.0 |
| Partial (low vision) | Regular | 0.0 | 1.0 |
| | Large | 1.0 | 0.0 |
| No | Regular | 0.0 | 1.0 |
| | Large | 0.0 | 1.0 |

Thus, we see that for reception demand to be met, characteristics of the person (i.e., see) must be properly matched to features of the task performance situation (i.e., font size). The assessment planner has little or no control over the person characteristics (e.g., see) and so must focus their efforts to provide task features that will allow reception demand to be met for individuals with diverse levels of sight (yes, partial, no).[50] (Obviously, if see = no, then large font will not allow reception demand to be met and a different task feature would need to be made available.)

### *Focal and Nonfocal Knowledge, Skills, and Abilities*

Consistent with what one would expect in a test of reading comprehension, sight is not a focal KSA but, rather, a nonfocal KSA; indeed there is a nonfocal requirement for sight (when a test is administered under default conditions).[51] In this case, the fact that see is not part of the target of measurement is fairly obvious. In other cases, whether a particular KSA is a nonfocal or focal KSA is not as obvious.

### *Meet Reception Demand*

In Table 12 we see conditional probabilities for meet reception demand.

Notice that these conditional probabilities are the machinery that embeds the warrant in a probability-based assessment. Each row of Table 12 may be seen as a rule or generalization that is part of the warrant of the assessment argument. Obviously, we should be using theory and

experience to provide the best backing for these warrants that we can.[52] As suggested in the table, the two rules pertaining to a test taker with low vision are as follows:

- IF see = partial (low vision) AND font size = regular THEN meet reception demand = no.
- IF see = partial (low vision) AND font size = large THEN meet reception demand = yes.

### *Three Nodes: Kind of Item, Reading Comprehension Demand, and Meet Reading Comprehension Demand*

Now let us add three additional nodes directly related to the demand for reading comprehension ability. These are kind of item, RC demand, and meet RC demand.

The kind of item node pertains to another feature of the task situation. We have specified two values for kind of item—main idea and find-the-word. A main idea item asks the test taker to identify the main idea of a reading passage. On the other hand, a find-the-word item asks the test to find a specified word in the passage.

The kind of item node is the parent node of the RC demand node. Indeed, in our simple model, RC demand is modeled as being entirely dependent on or driven by the value of kind of item. Specifically, if kind of item is main idea then RC demand is significant whereas if kind of item is find-the-word then RC demand is negligible.[53] This is sensible since it arguably takes very little (if any) reading comprehension ability to find a specified word in a passage while determining the main idea is likely to demand a significant amount of reading comprehension thinking from the test taker.

The meet RC demand node indicates whether the focal requirement for RC ability has been met—much as meet reception demand indicates whether the nonfocal requirement for reception has been met. The meet RC demand node has two parents, RC and RC demand. We can see that of the four combinations of the parent nodes, three would result in RC demand being met. For example, looking on the last row of the table, even if a person has poor reading comprehension ability (RC = poor) but the RC demand of the task is negligible (such as would arise from using a find-the-word item), the demand for RC ability is met (i.e., meet RC demand = yes).[54]

**Table 13**

***Meet Reading Comprehension Demand Depends on Reading Comprehension Demand and Reading Comprehension***

| Parent node(s) | | Meet RC demand | |
|---|---|---|---|
| RC demand | RC | Yes | No |
| Significant | Good | 1.0 | 0.0 |
| | Poor | 0.0 | 1.0 |
| Neglible | Good | 1.0 | 0.0 |
| | Poor | 1.0 | 0.0 |

*Note.* RC = reading comprehension.

As shown in Table 14, the only way to have good effective RC is to have both reception demand and RC demand met (satisfied).

**Table 14**

***Effective Reading Comprehension Depends on Meet Reception Demand and Meet Reading Comprehension Demand***

| Parent node(s) | | Effective RC | |
|---|---|---|---|
| Meet reception demand | Meet RC demand | Good | Poor |
| Yes | Yes | 1.0 | 0.0 |
| | No | 0.0 | 1.0 |
| No | Yes | 0.0 | 1.0 |
| | No | 0.0 | 1.0 |

*Note.* RC = reading comprehension.

Having added these various nodes to the model, let us now rearrange them in a way that highlights their relationships to some key models of the ECD CAF, as shown in Figure 11.[55] This rearrangement of the display of the Bayes net model does not affect its functioning. Within the model that we have built, the student model is represented by effective RC, the evidence model is represented by the item score, and the task model is represented by kind of item and font size.

The Bayes net in Figure 11 shows the case in which the person has low vision (see = partial) and good RC. She receives a test using a main-idea item (kind of word = main idea) and receives it using regular sized font (font size = regular). Because the individual's level of sight is unable to satisfy the nonfocal requirement for sight generated by the use of regular-sized font, reception demand is not met (meet reception demand = no) and her effective RC is poor, despite

the fact that her RC ability is sufficient to satisfy the focal requirement for RC ability (as evidenced by the node that says meet RC demand = yes). There is an 80% chance of her answering the item incorrectly. (Guessing on five-option multiple choice item accounts for the 20% probability of answering correctly.)

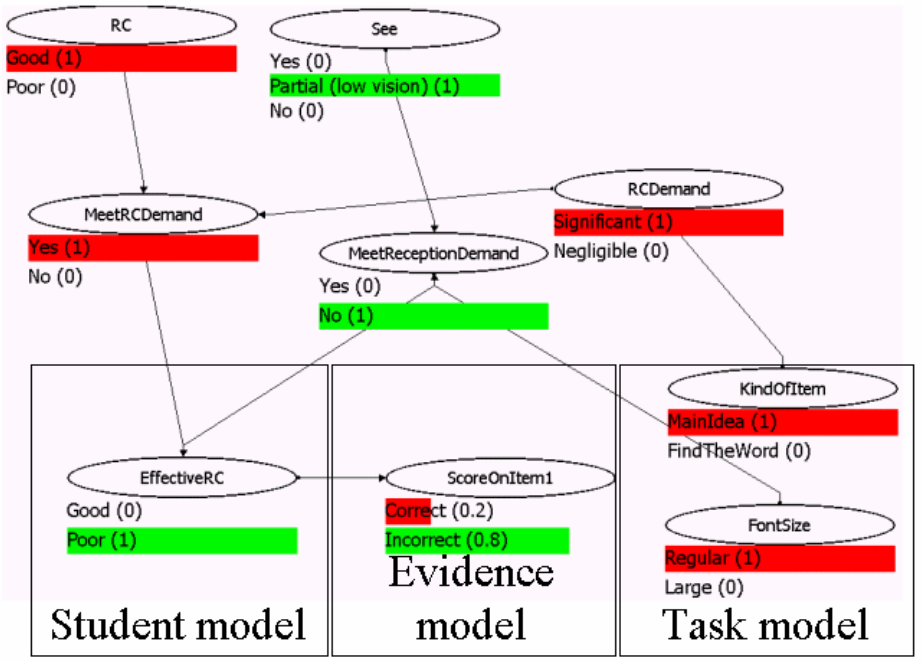


***Figure 11.*** **Bayes net, rearranged to showing relationship to key models of the conceptual assessment framework.**

*Note.* This portrays a false-negative outcome, based on mismatch between the test taker's vision and the font size.

Note that in the context of this report, one can pay close attention to nodes of targeted proficiency, effective proficiency, and meet reception demand but pay little attention to the scores themselves. This link becomes important for later stages of reasoning, both to account for measurement error in inferences about individuals and to guide test assembly so as to minimize measurement error in various designs for accommodated tests.

### *Another Definition of Validity*

The addition of new nodes provides a richer representation of assessments, but has new possible sources of uncertainty and, hence, new alternative explanations. Additional alternative

explanations invite us to consider an additional possible definition for validity. This definition has three criteria.

1. The outcome of comparing RC and effective RC is true (true-positive or true-negative).
2. Reception demand is met.
3. The requirement for the focal KSA (RC) is maintained at the same level faced in a known case of valid measurement (e.g., nondisabled examinees under default conditions).

The third criterion, which is new, presumes (in keeping with assumptions discussed later in this report) that demand for the targeted proficiency for nondisabled test takers under default conditions is appropriate, as would be the case when the kind of item is main idea as opposed to find-the-word. Thus, the third criterion would avoid situations such as would exist if a person with poor RC has good effective RC because the kind of item is improperly set to find-the-word, as shown in Case 4 below. This case may seem trivial, since this rather obvious case would already have failed on Criterion 1 (since it does not provide a true outcome).[56]

Yet Criterion 3 also addresses a subtler validity issue that could arise even if both Criteria 1 and 2 are satisfied. Consider a situation in which a person with low vision and *good* RC ability receives the find-the-word item in large font. Criterion 1 is satisfied because the outcome is true (i.e., true-positive: both RC and effective RC are good). Criterion 2 is also satisfied because the large font is appropriate for low vision and the person can, therefore, receive the test content. However, validity has quite arguably been compromised, since this individual faced only a reduced demand for RC ability due to the find-the-word item instead of the more appropriate main-idea item. Thus, according to Criterion 3, any deviation from default testing conditions (e.g., an accommodation) ought not reduce a demand for the targeted proficiency (i.e., kind of item must be main idea).[57] Essentially, Criterion 1 addresses the notion of basic fidelity to intent (at least in terms of outcome), Criterion 2 addresses the notion of accessibility (thereby reducing the likelihood for unfair disadvantages), and Criterion 3 addresses both unfair advantage to the test taker which could result from insufficient focal requirements and unfair *disadvantage* that could result from excessive focal requirements.

Before proceeding, we should note some important points about assessment that Figure 11 highlights. Note that the CAF models, which are blueprints for the operational elements of the assessment, formally incorporate only a few of the many variables that are essential in the assessment argument. Specification of CAF elements is necessary for coherent reasoning from assessment data, but the rationale for why their particular forms ground the targeted inferences about students is not part of this layer. This is why discussions about accommodations that focus on the elements themselves rather than the argument can prove unsatisfactory, and why extending the discussion to the assessment argument can clarify (if not conclusively answer) the questions that arise. Note also that the edge (arrow) connecting effective RC with the item score, in the student- and evidence-model boxes, represents the psychometric model employed in the assessment. It becomes clear how much of the assessment argument is presumed by considerations that lie outside the psychometric model, and why therefore simply manipulating psychometric models cannot provide a full understanding of the impact of accommodations in an assessment.

We will now look at four examples, one for each of four different outcomes—true-positive, true-negative, false-negative, and false-positive. For ease of comparison, all four cases will involve individuals with low vision (see = partial). For simplicity we will emphasize the simplest and least stringent index of validity—which merely involves comparing effective RC with RC—yet we will also note how these cases rate on the other validity criteria as well.[58]

Cases 1 and 2 represent the ideal situations for testing individuals with low vision. The font size is large, which allows the reception demand to be met. Furthermore, the kind of item is main idea, which causes a significant demand for RC ability.

In Case 1, the individual has good reading comprehension ability (RC = good) and performs well (effective RC = good) so the outcome is true-positive.

In Case 2, the individual has poor reading comprehension ability (RC = poor) and performs poorly (effective RC = poor) so that the outcome is true-negative.[59]

In Cases 3 and 4 there are problems which compromise validity. In Case 3, the individual with low vision is using a regular-sized font so the reception demand is not met and the performance (effective RC) is poor, even though the individual's ability (RC) is good. This is the situation that we originally had with Sue.

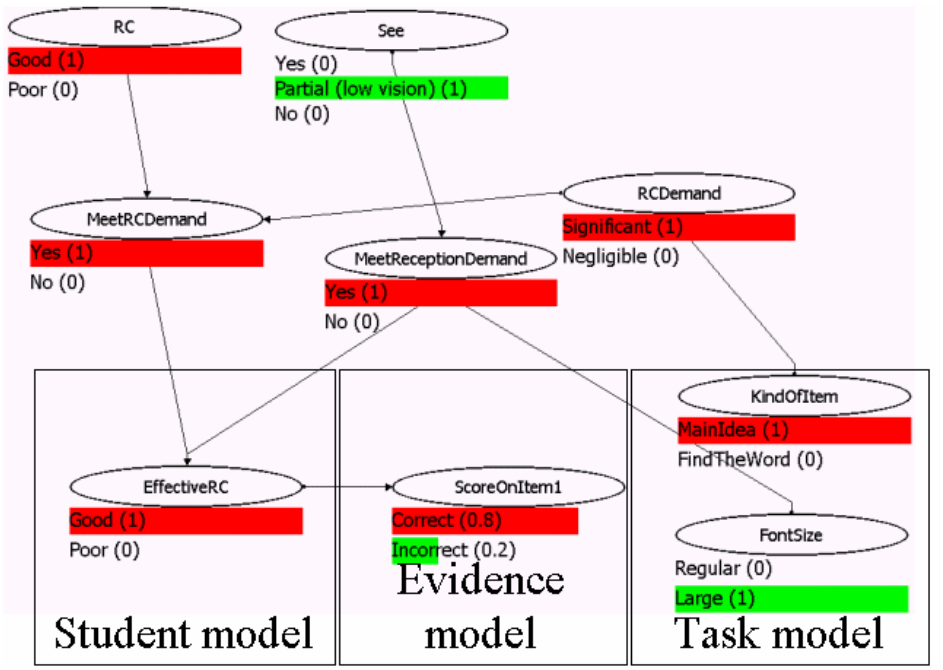


***Figure 12.*** **Case 1: True-positive, reception demand met, no reduction in reading comprehension demand.**

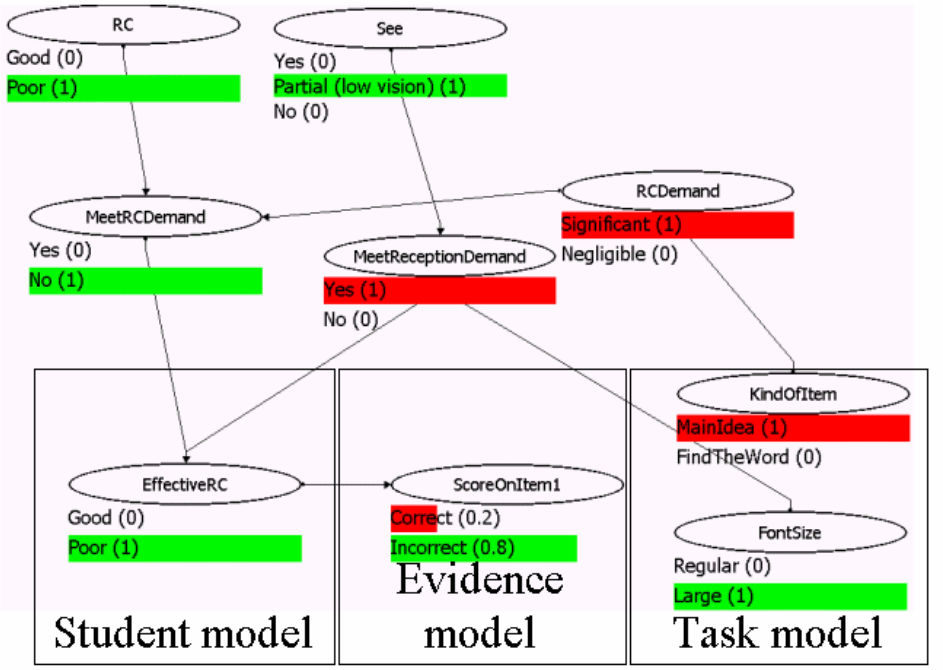


***Figure 13.*** **Case 2: True-negative, reception demand met, no reduction in reading comprehension demand.**

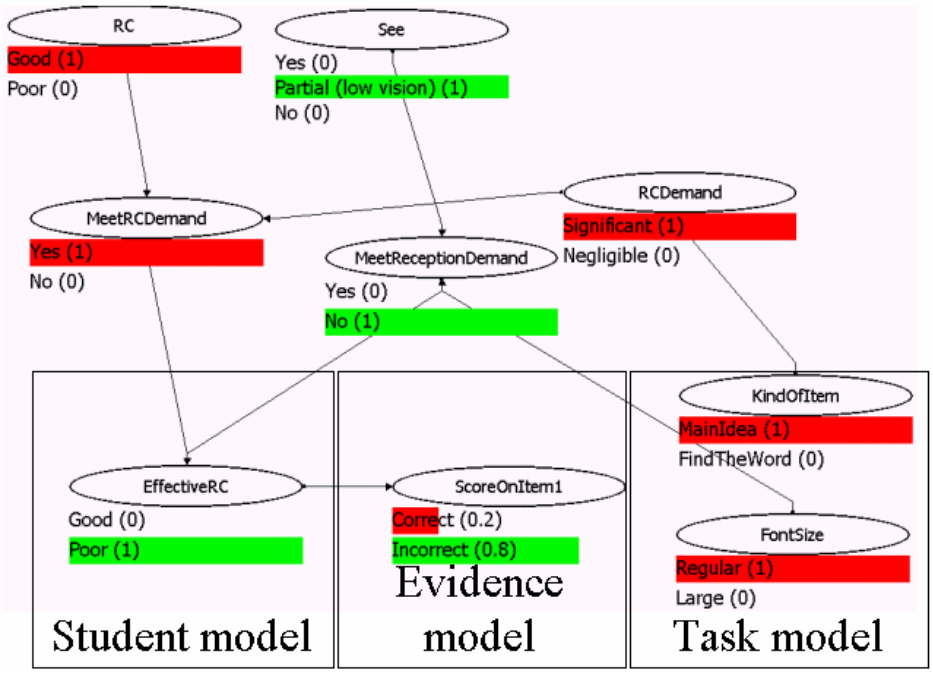


***Figure 14.*** **Case 3: False-negative, reception demand not met, no reduction in reading comprehension demand.**

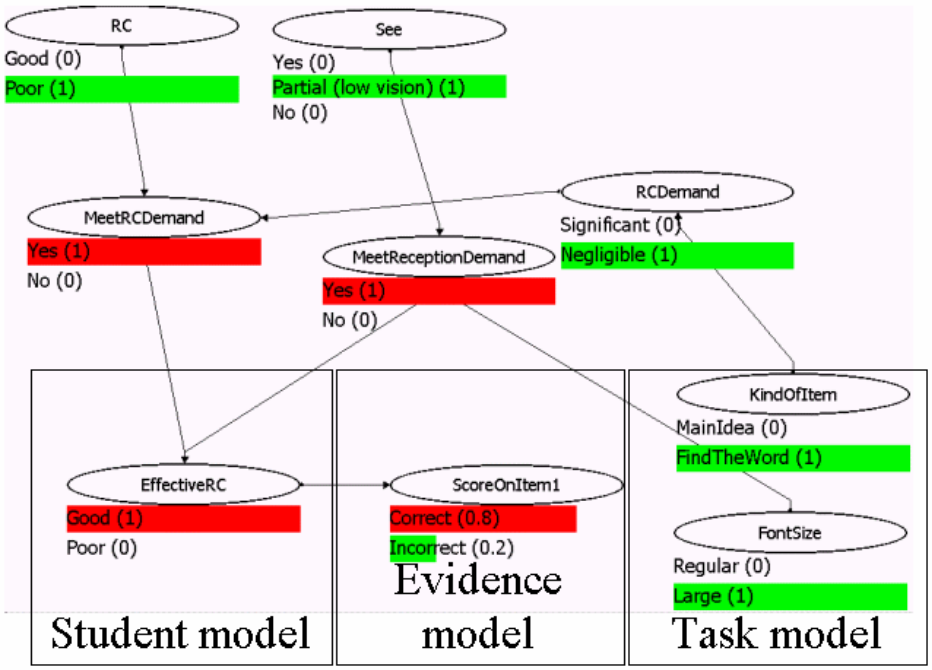


***Figure 15.*** **Case 4: False-positive, reception demand met, reduction in reading comprehension demand.**

In Case 4 the individual with low vision is using a large font so that their reception demand is being met and their performance (effective RC) is good. Yet this is a false-positive outcome because their actual ability (RC) is poor. Why is the individual able to perform well on the task despite poor ability? As one can ascertain from the graphic, the kind of item was improperly set to find-the-word, which meant that the item did not generate significant demand for reading comprehension ability. The individual was thus able to answer the item correctly despite their poor ability.

***Discussion of the Four Cases***

By modeling the assessment argument in this way, one can begin to anticipate situations that would yield good (valid) or bad (invalid) measurement. The Bayes net that we have been examining can be used to represent 24 situations. The number 24 can be derived from multiplying the two states of the focal KSA (RC, which is the only focal KSA in the targeted proficiency) times three learner states with respect to nonfocal KSAs times four task model variants as described below.

- 2 states of the focal KSA (RC)
    - good
    - poor
- 3 learner states with respect to the nonfocal KSA (see)
    - yes (unimpaired vision)
    - partial (low vision)
    - no (blind)
- 4 task model variants (2 x 2)
    - font size (regular, large)
    - kind of item (main idea, find-the-word)

Of these 24 situations, 15 satisfy Criterion 1 (true outcomes), 12 satisfy Criterion 2 (meet reception demand = yes), and 12 satisfy Criterion 3 (kind of item = main idea).[60] Only 6 of the

situations satisfy all 3 criteria, thereby satisfying the definition of validity discussed in section 3.10.

All six of these situations involve using a main idea kind of item. These 6 situations can be organized into three ability pairs, where each ability pair has one situation where the individual's RC ability is good and one situation where RC ability is poor. With respect to sight and font size the situation for the three ability pairs are as follows:

1. See = yes, font size = regular
2. See = yes, font size = large[61]
3. See = partial (low vision), font size = large

Ability Pair 1 represents the default testing conditions (font size = regular) for nondisabled test takers (see = yes). Ability Pair 3 represents the accommodated situation (font size = large) for individuals with low vision (see = partial). Ability Pair 2 represents the situation for nondisabled individuals (see = yes) who could potentially be appropriately tested using large font. The presence of Ability Pair 2 in this list suggest that large font might be considered as a possible universal design feature that could be made available generally (e.g., to persons with low vision or who are nondisabled) to individuals who desire it.[62]

One can argue that this is a lot of work to go to just to analyze situations and ascertain information that is intuitively obvious. Yet if this analysis can illuminate principles that help address situations that are more complex, then it is worth the effort. We believe that the current work lays groundwork for illuminating general principles and developing rigorous and systematic ways of developing runnable models of situations that are far more complex than what can be processed, stored, and shared using other techniques or by the human mind alone.

***Goals and Assumptions of the Model***

This section describes some of the goals and assumptions underlying the model. This section may not be of interest to all readers.

The distinction between RC and effective RC allows us to distinguish between two different kinds of error, which we call *specification error* (a term used in structural equation modeling and econometrics) and *measurement error* (a term commonly used in educational measurement).

We use the term measurement error to refer to errors that can be essentially eliminated given a sufficient number of observations (e.g., item scores). Examples of possible sources of measurement error include essentially random events such as careless mistakes. We model measurement error as pertaining strictly to the relationship between effective RC and the item score. Specification error, on the other hand, encompasses other sources of error (i.e., errors that could still exist even in the absence of measurement error, and that pertain to the relationship between RC and effective RC).[63] Thus, specification error is essentially about validity while measurement error is about reliability.

In the context of this discussion, we assume that the target proficiency is correctly defined and that the definition reflects, among other things, a deep understanding of the demands for focal KSAs in criterion situations. We assume that the only possible source(s) of specification error are those shown in the model (i.e., nodes that are the direct or indirect parents of effective RC).[64] In keeping with these assumptions, we assume that various features of the student and evidence models (notably item parameters; proficiency scales; cut-points, as applicable; and conditional probabilities for item scores) were formulated based on only test takers for whom the targeted proficiency could be accurately measured given available conditions (standard or accommodated).[65] This would require that individuals without disabilities be appropriately and accurately assessed under default testing conditions. It would also involve approval and provision of appropriate and successful testing accommodations, wherever applicable.[66] This would also involve successful and effective use of practice and familiarization materials as well as good test security and no cheating.

Unless otherwise stated, the model assumes a hypothetical full and constant population with a variety of possible knowledge or skill attributes. The same mild prior probability distribution posited for the distribution of ability in focal KSAs is, in this case, assumed to be the same for disabled and nondisabled individuals, which would be of negligible importance in applications in which enough data is gathered from each student to ground inferences about them individually.[67] Furthermore, our usage of the model typically involves scenarios in which focal and nonfocal KSAs are set to specific values so, from that standpoint these assumptions are not critical.

This model is intended to facilitate what-if reasoning that allows us to evaluate consequences of different choices in test design and test use, including not only optimal

assessment design and implementation choices (such as those that conform well to the assumptions above), but also a range of suboptimal choices. For example, the model is capable of representing inappropriate testing accommodations. Indeed we exploit this feature to evaluate the appropriateness of a potential testing accommodation. Furthermore, the model is capable of representing the capabilities of individuals who would be ineligible to take the test according to the assumptions above, because there is no available set of testing conditions (e.g., accommodations) that would allow good measurement to occur.

While not a specific assumption of the model, it is the intent of this report that users of the modeling techniques strive to include a wide range of accessibility-related task features that will permit fair and accurate measurement for individuals with diverse disabilities. This report also seeks to move toward a more comprehensive and unified approach that encompasses diverse strategies for meaningful participation of the widest range of individuals with disabilities, including but not limited to testing accommodations; universal design; using practice and familiarization materials; appropriate coaching; changes to eligibility rules for participation in the test; and, as appropriate, revision to the definition of the targeted proficiency, and so on.[68]

In summary, the model represents something of a normative or idealized situation that invites comparison against actual test settings and data. While no actual assessment design or operational test can fully meet these assumptions, we believe that many would come close enough to derive utility from the models discussed in this report. It should be noted that generally, the modeling activity has focused on accessibility and measurement related issues rather than logistical and cost-related issues, which necessarily also play a role in accommodation-related policies.[69]

### *Systematic Steps for Using the Approach to Promote Inclusion*

It is useful to pause a moment to consider a systematic approach for building and using validity argument models to promote inclusion of individuals with disabilities in assessments. These steps need not always be done in a specific order and some steps may be repeated. Although this report touches upon all steps listed below, it emphasizes Step 3, construction of the model.

1. Identify possible areas for improvement of inclusion and accessibility
2. Define what basic indices of validity one will use

3. Construct the model of the validity argument, with emphasis on parts related to disability.
4. Run the model
5. Verify the results then refine and redo Steps 1 through 4, if necessary
6. Apply the results

Step 3 includes substeps such as (a) Define relevant profiles of individuals of interest. Give consideration to characteristics related to disability (especially reading-related skills) and language status. Among the set of relevant profiles, it is suggested that one include the profile of an individual without any disability as well as individuals with diverse disabilities; (b) Identify the task performance situations, including both the default situation and deviations from it (e.g., accommodations). Focus on features that most directly affect demand for examinee KSAs. (c) Identify the key focal KSAs and nonfocal KSAs. Distinguish between KSAs that are essential components of the targeted proficiency (focal KSAs) and those that are not (nonfocal KSAs); (d)Further define the targeted proficiency. Identify more specifically the levels of focal KSAs that constitute having a good or successful level in the targeted proficiency. (e) Define effective proficiency using a range of situations selected earlier. Define how effective proficiency (essentially, performance under actual assessment conditions) is affected by the different combinations of person profiles (sets of characteristics) and task performance situations. Include consideration of individuals without disabilities taking the assessment under default conditions.

It is hoped that the reader will be alert to the steps as they are mentioned in this report, even if they do not occur in particular order indicated above. These steps are described in additional detail in the Discussion and Conclusions section.

### A More Complex Example: Blind and Read-Aloud

Let us consider another example. We need to measure the reading comprehension ability of Tim. Tim is blind and he proposes using a read-aloud accommodation, that is, having the test content read aloud to him by human reader or presented as synthesized or prerecorded speech. Tim is familiar with read-aloud formats (e.g., live reader, prerecorded speech, computer-generated speech), having previously used them on various literacy tests and no technological or

logistical barriers prevent such an accommodation. How should we respond to Tim's request? This kind of question arises in many large-scale assessments, such as NAEP.

### *NAEP Reading and the Read-Aloud Accommodation*

To respond to this request, we need to examine what the test is intended to measure. As noted earlier, NAEP allows the read-aloud in the mathematics assessment and some other assessments but not in the reading assessment. According to the 2003 NAEP reading framework, "because NAEP is a reading comprehension assessment, test administrators are not allowed to read the passages and questions aloud to students" (2003 NAEP Reading Framework, p. 3). However, the framework document does not detail specifically why a reading test should not be read aloud to the test taker. One possibility is that the ability to decode (i.e., to form words from characters) is part of what the assessment is intended to measure.[70] If decoding is a part of the proficiency targeted for measurement by the assessment, then providing a read-aloud accommodation would reduce demand for RC ability, since the read-aloud accommodation almost always involves the reader speaking whole words rather that speaking one letter at a time. Thus, providing a read-aloud accommodation would tend to bestow an unfair advantage on the person receiving the accommodation.

Is the NAEP reading assessment explicitly intended to include decoding as part of reading proficiency? The 2003 NAEP reading framework does not seem to state a position on this issue. However, the framework does mention decoding in the context of what would be expected of a diagnostic test of fourth-grade reading as opposed to a test of reading achievement, such as NAEP:

> The assessment does not focus solely on the many specific skills a reader must use but seldom uses in isolation. This is in keeping with NAEP's role as an assessment of overall achievement rather than a diagnostic tool for individual students. . . .
>
> The NAEP reading assessment is an assessment of overall achievement, not a tool for diagnosing the needs of individuals or groups of students. A diagnostic assessment of reading in grade 4 would examine an individual student's ability to read fluently aloud, using both the ability to *decode* words and to recognize them instantly. It would explore what specific comprehension skills the reader could demonstrate such as finding the main idea, relating cause and effect, inferring character qualities, and detecting sequence.

> However, an achievement measure such as NAEP asks broader questions: for example, how well does this student or group of students read? Is this level of achievement good enough to meet the standard that has been set?
>
> NAEP examines whether students can use multiple skills, not specific skills, to comprehend what they read. Effective reading programs definitely focus on teaching specific reading skills. However, when people actually read, they choose and orchestrate arrays of skills, sometimes almost simultaneously. The NAEP reading assessment examines whether students can actually use sets of skills in reading for different purposes. (pp. 35–36, emphasis added)

We can summarize as follows:

1. It is not clear whether decoding is an intended part of NAEP reading proficiency. The framework is not explicit in this regard.
2. The accommodation policy against the read-aloud accommodation seems to suggest that decoding might be part of what the NAEP reading assessment is intended to measure.
3. Decoding is part of what the NAEP assessment planners believe a diagnostic assessment of reading ability would include, at least at grade 4.
4. While decoding may be part of the multiple specific skills that NAEP believes constitute good reading, NAEP does not attempt to diagnose these specific skills.

Let us examine the situation in which decoding is part of what the assessment is intended to measure, giving us the basic situation shown in Figure 16. Let us call this *Definition A* of reading comprehension.

This basic analysis suggests that in this situation, providing a read-aloud accommodation would provide an unfair advantage to the person receiving the accommodation and that, therefore, the accommodation should not be allowed. Where the decoding is not part of the target of measurement—let us call this *Definition B*—a read-aloud accommodation does not yield a credible alternative explanation and, therefore, we expect the read-aloud accommodation to yield valid inferences. In other words, where there are no credible alternative explanations for the

observations (scores), then the intended explanation—that they were caused by the person's unseen level of reading comprehension—proceed normally and at full strength (see Figure 17).

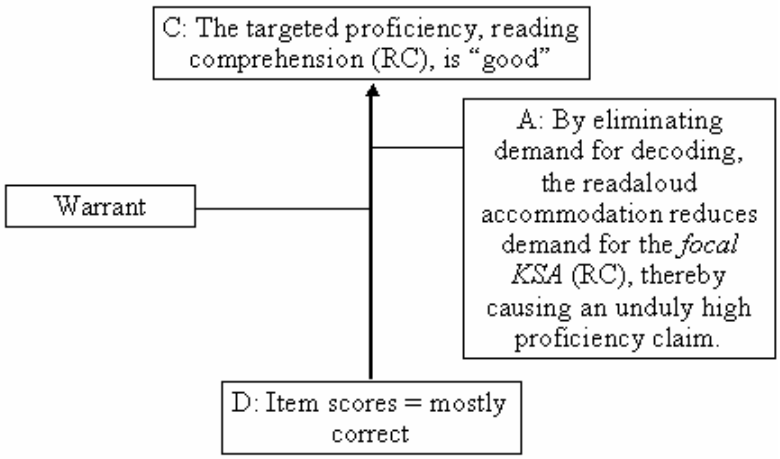


***Figure 16.*** **Where decoding is part of the target of measurement (Definition A), a read-aloud accommodation yields a credible alternative explanation.**

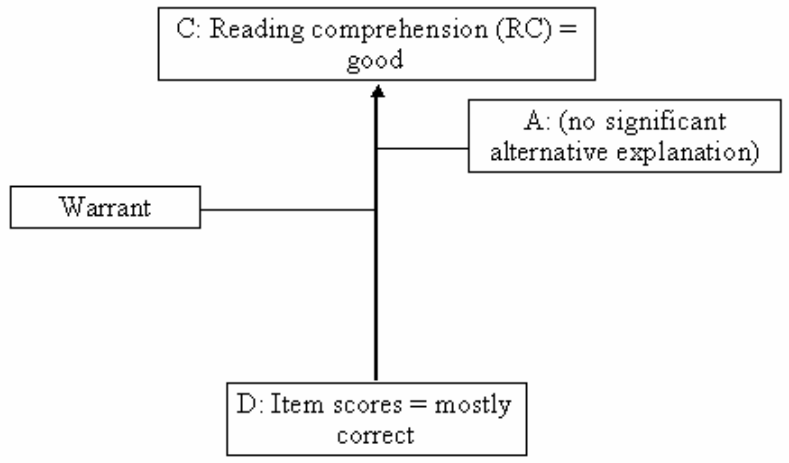


***Figure 17.*** **Where decoding is not part of the target of measurement (Definition B), a read-aloud accommodation does not yield a credible alternative explanation.**

Focusing on the option of providing a read-aloud accommodation for Tim, we see, again, that the Toulmin diagram was useful in framing the basic issue. Yet this structure is of limited value in thinking through the details of an accommodation decision, even if the issue of whether decoding is part of reading has been clearly stated. For example, following are a few considerations:

- Could Tim take the test via braille? Wouldn't a braille accommodation still require decoding ability and, therefore, largely bypass the decoding issue?[71]
- How important is decoding as a component relative to other key components of reading comprehension—such as the comprehension component? If decoding is a relatively minor constituent of reading comprehension, then perhaps providing the accommodation would not seriously compromise validity.
- Is the actual level of decoding required to do well on the operational items consistent with the intended role of decoding as specified in the definition of the construct? For example, is it possible that the actual decoding demand of items is higher than it should be?
- What would be the implications of changing the definition of the targeted proficiency (reading comprehension)?[72]
- Are there alternative forms of evidence that would be more appropriate for this individual and for the decisions that need to be made?
- What is typical or common practice among large-scale assessments?

The practical necessity of dealing with such complexities spurs us to find richer representations of the situation.

The Bayes net we developed for the situation of Sue and low vision would provide a useful foundation, but is has several shortcomings. For example, it did not provide task features that would be helpful to individuals who are blind (e.g., braille or read-aloud presentation formats). It did not break reading comprehension (RC) down into component skills (e.g., comprehension and decoding) nor did it provide a way of representing the impact of demands for those skills on effective RC.[73]

### *A General Schema*

What would a richer representation look like? To a large extent, it might simply be an elaboration of the general approach we used for Sue and low vision. Figure 18 shows a schematic representation of that pattern.

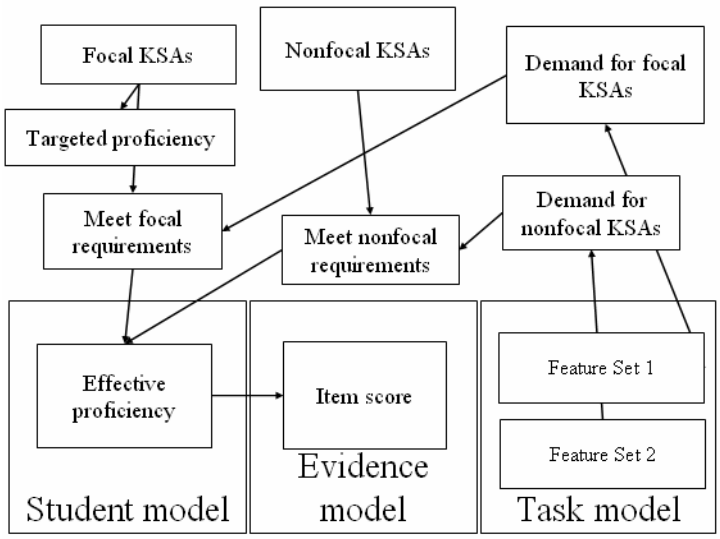


***Figure 18.*** **A general schema.**

Following are some key points regarding the general schema:

- Targeted proficiency is the intent of measurement. It is composed of one or more focal KSAs.[74]
- Focal KSAs and nonfocal KSAs describe characteristics of the test taker.
- The term nonfocal KSA (or its synonym, nonfocal skill) is agnostic as to whether good effective proficiency depends on it, so we use the term nonfocal requirement to denote that good effective proficiency does depend on it. For most academic assessments, the sense of sight is a nonfocal requirement.
- Effective proficiency is one of many possible instances of the targeted proficiency as operationalized under specific conditions.
- Where there are unmet nonfocal requirements, effective proficiency tends to be unduly poor.
- Where demands for focal KSAs have been reduced, effective proficiency tends to be unduly good.
- The item scores are modeled as being driven by effective proficiency plus measurement error.

- Possible criteria for validity might include (a) fidelity to intent (derived from comparing targeted proficiency to effective proficiency), (b) satisfaction of nonfocal requirements, (c) reception demand being met (if different than b), and (d) maintenance of focal requirements relative to a known case of valid measurement, and so on.
- Accommodations generally seek to revise, reduce, or eliminate demands for nonfocal KSAs in which there is a deficit and rely instead on nonfocal KSAs where there is no deficit.
- Unfair advantage will tend to result if features intended to overcome accessibility barriers (by changing demands for nonfocal KSAs) also inadvertently reduce demand for the focal KSAs comprising the targeted proficiency.[75]
- It is sometimes advantageous to bypass some of the nodes in the creation of a model, especially where a node would have only one parent. For example, it is common to bypass nodes for demand for a focal KSA or demand for a nonfocal KSA.[76]
- Generally, the more an actual situation deviates from the assumptions underlying the model, the more complicated the argument.

Figure 19 applies this general schema to the situation of Sue and low vision. Reception demand is not met, leading to a false-negative outcome (RC = good, effective RC = poor).

Figure 20 applies this general schema to the situation of Sue and low vision after proper accommodation. Reception demand is met, leading to a true-positive outcome (RC = good, effective RC = good).

### *A Model for the Situation of Tim and Read-Aloud*

Utilizing the general schema, we can develop a simple representation of the argument for Tim and read-aloud. As shown in Figure 21, under Definition A (decoding is part of reading comprehension) and the read-aloud accommodation, the outcome is false-positive. Tim's targeted proficiency (RC) is poor, due to poor decoding ability, yet his effective proficiency is good. Notice the thick arrow leading from presentation mode = read-aloud to the demand for focal KSAs node. This thick arrow is outside what is part of the basic template for the general schema. This represents the idea that while the read-aloud presentation mode was not intended to

affect demand for focal KSAs (or the targeted proficiency that they constitute), such an effect nevertheless occurs. (Normally, the purpose of the presentation mode accommodation would be to reduce demand for nonfocal skills in which there was a deficit and to instead rely on nonfocal skills in which there were no deficits.) Specifically, the read-aloud reduces demand for the focal KSA of decoding, thereby allowing a person with poor decoding and hence poor reading comprehension proficiency to perform well on the item, thus resulting in a false-positive outcome.

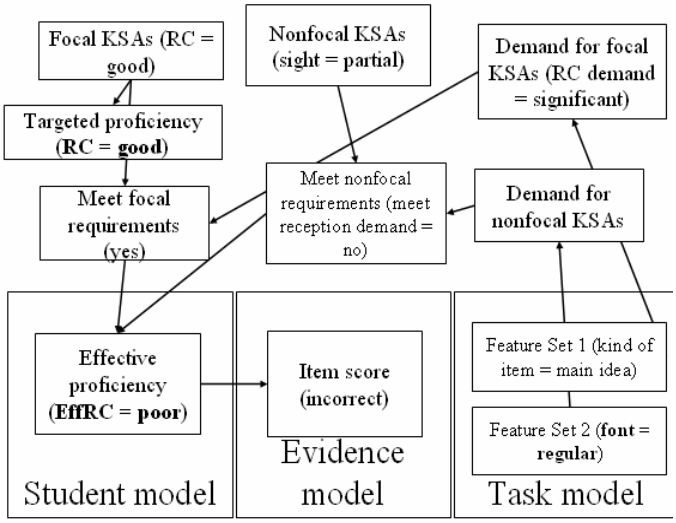


***Figure 19.*** **The original situation of Sue and low vision.**

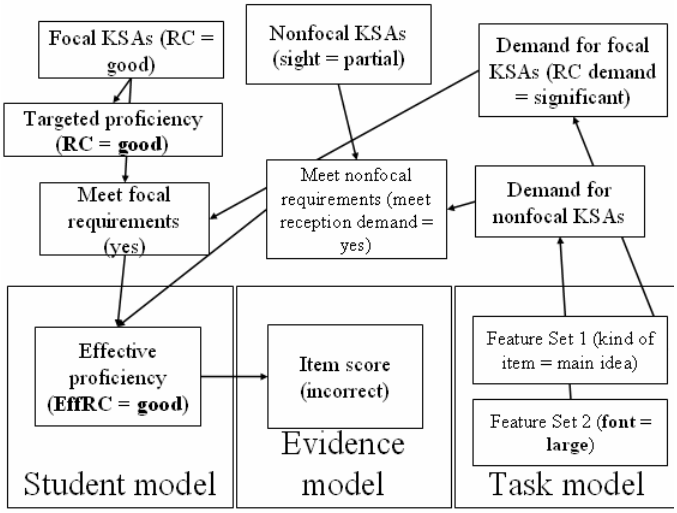


***Figure 20.*** **Sue and low vision, with proper accommodation.**

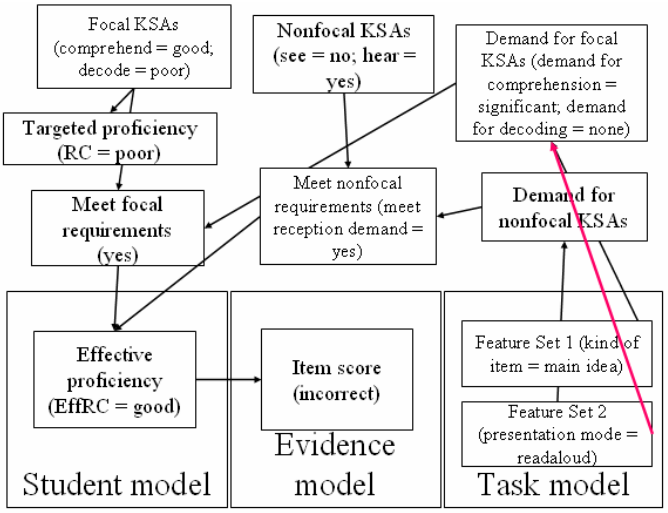


***Figure 21.*** **Tim and read-aloud under Definition A.**

As shown in Figure 22, under Definition B (decoding is not part of reading comprehension), under exactly the same task performance conditions, the outcome is true-positive (both effective proficiency and targeted proficiency are good).

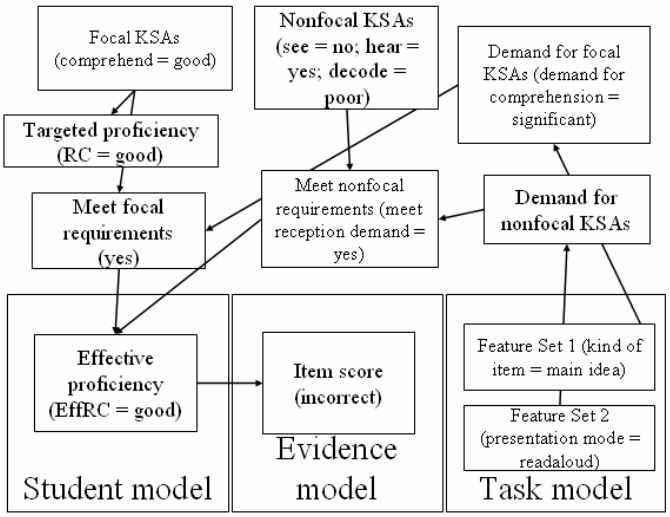


***Figure 22.*** **Tim and read-aloud under Definition B.**

Despite the advantages of the general schema, it has some limitations, among them the fact that it hides much of the detail of the interactions among variables. Another limitation is that the general schema, in its simplest form, may not represent all connections among categories of variables. For example, as in Figure 21, a task feature that was intended merely to affect

nonfocal requirements (the read-aloud accommodation intended to eliminate reliance on sight) also inadvertently reduces a focal requirement (decoding ability), a situation that involves connections beyond the simplest versions of the general schema.

### *A Richer Bayes Net Representation*

Let us consider a yet richer representation of the situation of Tim. This Bayes net, shown in Figure 23, has 23 compared to 9 nodes for the main Bayes net for Sue and low vision. This Bayes net can help clarify how the definition of the targeted proficiency affects the validity argument.

Recall that, under Definition A, a person must have both good comprehension skill and okay (or better) decoding skill in order to have good reading comprehension. On the other hand, under Definition B, one need only have good comprehension skill (i.e., even the lowest level of decoding [poor] is adequate). These definitions are represented in Table 15.

**Table 15**

***Two Possible Definitions of Reading Comprehension: A and B***

| | Definition of reading comprehension | |
|---|---|---|
| KSA | A | B |
| 1. Comprehend (good, poor) | Good | Good |
| 2. Decode (good, okay, poor) | Okay | n/a |
| Decoding | Focal KSA | Nonfocal KSA |

*Note.* KSA = knowledge, skills, and abilities.

The way that these definitions are represented in the Bayes net is shown in Figure 23, which shows conditional probabilities encoded for the RC node of the Bayes net.

Thus, using this Bayes net one is able to evaluate how the appropriateness or validity of specific accommodations (e.g., read-aloud) is influenced by the definitions of the target of measurement (A versus B). The same Bayes net can be used to evaluate the impact of other assessment design changes, such as the influence of accommodations on nondisabled test takers.[77] Table 17 describes the variables (and their levels) in this Bayes net.

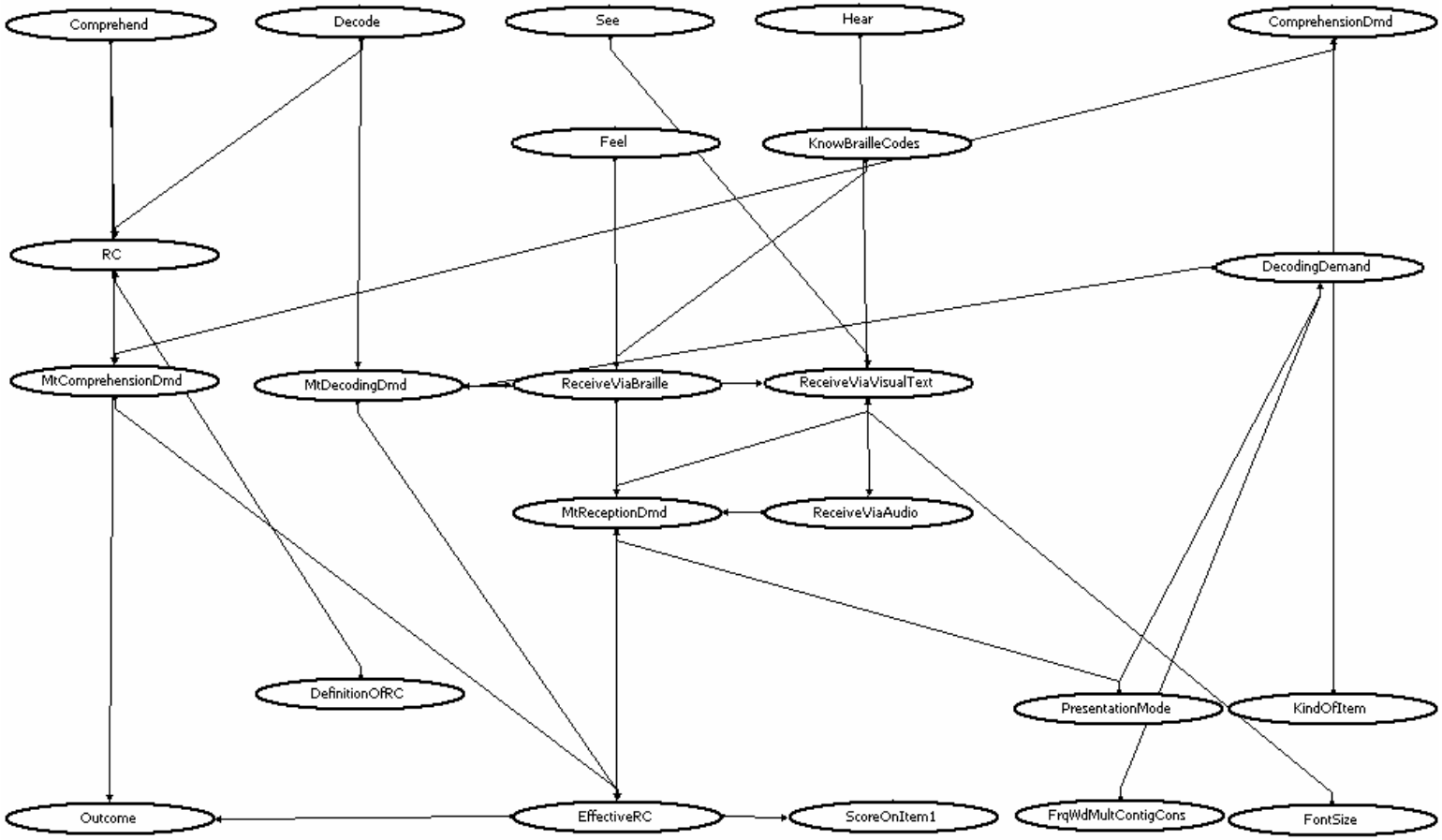


***Figure 23.*** **A picture of the Bayes net.**

**Table 16**

***Under Definition B It Is Possible to Have Poor Decoding Ability and Still Have Good Reading Comprehension Ability***

| Parent node(s) | | | RC | |
|---|---|---|---|---|
| Definition of RC | Comprehend | Decode | Good | Poor |
| A: Decoding relevant | Good | Good | 1.0 | 0.0 |
| | | Okay | 1.0 | 0.0 |
| | | Poor | 0.0 | 1.0 |
| | Poor | Good | 0.0 | 1.0 |
| | | Okay | 0.0 | 1.0 |
| | | Poor | 0.0 | 1.0 |
| B: Decoding irrelevant | Good | Good | 1.0 | 0.0 |
| | | Okay | 1.0 | 0.0 |
| | | Poor | 1.0 | 0.0 |
| | Poor | Good | 0.0 | 1.0 |
| | | Okay | 0.0 | 1.0 |
| | | Poor | 0.0 | 1.0 |

*Note.* See the row where definition of RC = B, comprehend = good, and decode = poor. Compare this with the similar situation but where definition of RC = A. RC = reading comprehension.

With this model, we can see whether good or bad measurement would result from a wider range of conditions. This model deals with disability statuses such as being deaf (hear = no), blind (see = no), dyslexic (decode = poor), and neuropathic (feel = no), and combinations thereof (e.g., deaf-blind), plus no disability. There are 36 task model variants (calculated by multiplying the number of levels of each of the four task model variables: 3 x 2 x 3 x 2 = 36), some of which involve accommodations (large font, read-aloud, braille). In addition there are two definitions of the targeted proficiency available (A and B).

Table 18 lists some of key variables (nodes) and lists states of variables for analyses based on two definitions of reading comprehension (Definition A and Definition B). The outcome node automates the comparison between RC and effective RC. The definition of RC node allows the model user to select between two definitions of RC—A and B. One can see from this table that there are only three rows for which the analyses differ—definition of the targeted proficiency, RC, and outcome.

**Table 17**

***Variables in the Richer Model***

| No. | Subcategory of variable | Variable | Levels |
|---|---|---|---|
| 1 | Focal KSA | Comprehend | Good, poor |
| 2 | Nonfocal KSA | See | Yes, partial, no |
| 3 | Nonfocal KSA | Hear | Yes, no |
| 4 | Focal KSA for Definition A, nonfocal KSA for Definition B | Decode | Good, okay, poor |
| 5 | Nonfocal KSA | Know braille codes | Yes, no |
| 6 | Nonfocal KSA | Feel | Yes, no |
| 7 | Task model variable | Presentation mode | Visual text, read-aloud, braille |
| 8 | Task model variable | Kind of item | Main idea, find-the-word |
| 9 | Task model variable | Frequency of words with multiple contiguous consonants | High, medium, low |
| 10 | Task model variable | Font size | Regular, large |
| 11 | Targeted proficiency | RC | Good, poor |
| 12 | Student model variable | Effective RC | Good, poor |
| 13 | Intervening variable | Meet comprehension demand | Yes, no |
| 14 | Intervening variable | Meet decoding demand | Yes, no |
| 15 | Intervening variable | Comprehension demand | High, low |
| 16 | Intervening variable | Decoding demand | High, medium, low |
| 17 | Intervening variable | Receive via visual text | Yes, no |
| 18 | Intervening variable | Receive via braille | Yes, no |

*(Table continues)*

Table 17 (continued)

| No. | Subcategory of variable | Variable | Levels |
|---|---|---|---|
| 19 | Intervening variable | Receive via audio | Yes, no |
| 20 | Intervening variable | Meet reception demand | Yes, no |
| 21 | Decision variable | Definition of RC | A, B |
| 22 | Output variable | Outcome | True-positive, true-negative, false-positive, false-negative |
| 23 | Evidence model variable | Score on item 1 | Correct, incorrect |

*Note.* KSA = knowledge, skills, and abilities, RC = reading comprehension.

As the table shows, using this model with Definition A, a blind individual with good comprehension and poor decoding ability who receives a read-aloud accommodation has a false-positive outcome. (This outcome is derived from the fact that while RC is poor, effective RC is good.) On the other hand, under Definition B, with all other inputs being the same as in the previous example of Definition A, the outcome is true-positive. (This can be ascertained from the noticing that RC and effective RC are both good.)

Notice that in running these analyses, we are essentially doing a what-if kind of analysis. We are treating certain variables (e.g., comprehend, decode) as if their values were known in order to carry out what-if analyses that illuminate implications for other variables in the network (e.g., outcome [true or false]).

These analyses demonstrate the importance of the definition of the targeted proficiency in determining whether an accommodation such as read-aloud is likely to yield valid inferences. This model of the argument for reading comprehension is able to represent a much wider array of situations than our earlier representations. Features and benefits of the richer Bayes net representations are elaborated on in Appendix C.

### *Role of Psycho-Physical Modeling*

One of the most important conceptual tools for increasing the usefulness as well as managing the complexity of the nets is what we term psycho-physical (i.e., psychological-

physical) models. These are essentially, models (or mini models) of the psychological and physiological processes underlying successful (and unsuccessful) performance. Such models might entail any of a variety of theoretical orientations (e.g., cognitive, information processing, physiological). Among other things, psycho-physical models connect the person with the assessment and its environment. For example, a psycho-physical model seeks to connect the largely invisible KSAs with the visible world of task features.[78] The critical points of juxtaposing the invisible with the visible are those nodes that determine whether task-generated demands for person characteristics can be satisfied.

**Table 18**

***Two Analyses Using the Model, Highlighting the Decode and Outcome Variables That Are Different Between the Analyses***

| Variable (node) | Input vs. output variable | Analysis A | Analysis B |
|---|---|---|---|
| 1. Definition of the targeted proficiency | Input | A | B |
| 2. Comprehend | Input | Good | Good |
| 3. Decode | Input | Poor | Poor |
| 4. RC | Output | Poor | Good |
| 5. Effective RC | Output | Good | Good |
| 6. Outcome (of comparing RC with effective RC) | Output | False-positive | True-positive |
| 7. See | Input | No | No |
| 8. Hear | Input | Yes | Yes |
| 9. Know braille codes | Input | No | No |
| 10. Feel | Input | Yes | Yes |
| 11. Presentation mode | Input | Read-aloud | Read-aloud |
| 12. Kind of item | Input | Main-idea | Main-idea |
| 13. Frequency of words with multiple contiguous consonants | Input | Medium | Medium |
| 14. Font size | Input | Regular | Regular |

*Note.* KSA = knowledge, skills, and abilities, RC = reading comprehension.

A good psycho-physical model reduces the complexity of the domain model for a given level of domain model sophistication. Consider, for example, the recent application involving reading comprehension. One could have modeled each of six person variables (e.g., comprehend, see, decode, hear, know braille codes, and feel) and four task model variables (font size, presentation mode, kind of item, frequency of words with multiple contiguous consonants)—all as direct parents of effective reading comprehension. However, this would have required the creator of the model to define (based on theory and experience) over 10,000 conditional probabilities—essentially the product of 11 numbers—the numbers of states in (a) each of the 10 person and task variables and (b) their 1 child variable.

Through the use of an appropriate psycho-physical model, we can make these variables indirect parents of effective reading comprehension and can simplify the situation. By using a psycho-physical model, we can establish intervening variables (that might be called collector variables), that are both (a) children of the person and task variables and (b) parents (either direct or distant) of effective reading comprehension.[79] One effect of the use of such psycho-physical model is that instead of needing to define over 10,000 conditional probabilities, less than 200 are actually required using intervening variables that reflect a psycho-physical model of performance. For some purposes, it may be valuable to use closer to 10,000 conditional probabilities rather than to 200, yet it seems fair to say that a good psycho-physical model can greatly simplify the building of Bayes net models. It also seems fair to say that with some psycho-physical model in mind, one is more likely to know what values are likely to be appropriate with any conditional probabilities that one must define.

At least as important as the ability to reduce the complexity of model creation, a good psycho-physical model will separate, to the extent feasible, psycho-physical processes that, for a given assessment, are completely or mostly construct-relevant (e.g., meet comprehension demand) from those that are completely or mostly construct-irrelevant (e.g., meet reception demand). Clear separation of the construct-relevant from the construct-irrelevant components tends to simplify the validity argument of the assessment.[80] A good psycho-physical model will tend to support the distinction made between focal and nonfocal KSAs. Furthermore, the specific nature of the psycho-physical model employed will influence which set of validity criteria is likely to be most useful in any particular setting.

## A Simple Model for NAEP Reading and Mathematics

Let us now examine another model that reflects some of the concerns and characteristics of NAEP reading and mathematics. We will begin by describing some considerations in creating a Bayes net model that attempts to address issues regarding some accommodations for NAEP reading and mathematics. We will then run the model for several basic illustrative cases, highlighting issues for further investigation. We will then have a brief discussion of the relationship between this modeling activity and empirical research. Finally, we then use the example of a dictionary feature—a feature that brings up additional issues.

This model attempts to address several accommodations. Among these are (a) one that is provided by NAEP for both reading and mathematics (large-print booklet, or large font); (b) one provided only for mathematics (read-aloud); and (c) one that is rarely, if ever, provided in either reading or mathematics (braille). The model also addresses the use of two varieties of English-language dictionary, which will also be briefly addressed in this report. To the best of our knowledge, neither of these dictionaries is now offered by NAEP to individuals with disabilities. Consideration of the dictionaries seems useful in pointing to potential future models of the bilingual dictionary and bilingual mathematics booklet accommodations for English-language learners (in NAEP mathematics only). (This report focuses on students with disabilities rather than English-language learners.)

### *Background on NAEP Accommodations*

Generally, NAEP attempts to provide accommodations required by individual education plans (IEPs). Table 19 shows the accommodations most frequently provided by NAEP. As shown below, the read-aloud accommodation is not allowed for reading. The bilingual accommodations (bilingual booklet and bilingual dictionary) apply to limited-English-proficient (LEP) students. Extended time is commonly used with other accommodations.

Some accommodations can be combined in the model, thereby constituting accommodation packages. For example, one could have a large-print booklet along with either one of the dictionaries (presumably also with large font, if needed).

The model discussed in this section makes a number of assumptions and simplifications. One of the simplifications is that this model does not distinguish between the three grade levels (4, 8, and 12) of NAEP reading and mathematics. Furthermore, it treats all items in a given

assessment as though they all make the same physical and cognitive demands upon the examinees (rather than modeling each item differently). For example, the NAEP frameworks has three different proficiency levels (basic, proficient, and advanced), plus additional score designations within each level; yet, the Bayes net implementation that we used provided only two levels for each proficiency (good, poor) and usually no more than three levels for the other variables. Some simplifications were also made where a more nuanced representation of the situation would be unmanageably complex for the purposes of this presentation. With respect to barriers to access, the model addresses accommodations related to presentation of assessment content rather than including issues such as timing or the ability to record answers.

**Table 19**

***Accommodations Most Frequently Provided by NAEP***

| Accommodation | Comment |
|---|---|
| Bilingual booklet | Mathematics only |
| Bilingual dictionary[81] | Mathematics only |
| Large-print booklet | |
| Extended time | |
| Read-aloud | Not permitted in reading |
| Small group administration | Usually in connection with read-aloud |
| One-on-one administration | Usually in connection with read-aloud |
| Scribe or use of computer | |
| Other | |

The ECD framework is in several respects a richer and more comprehensive framework and contains additional structure that is absent from the NAEP framework.[82] Hence, it was necessary to make guesses or assumptions to fill in some of the gaps. Despite these simplifications and limitations, we believe that the present discussion can show ways of making more explicit the validity argument for NAEP and other large-scale assessments.

### *Challenge of Mapping From the Framework Documents Into ECD*

Some difficulties were encountered in attempting to map information from the NAEP framework into the ECD validity argument framework. We found no explicit statement in the NAEP reading framework to indicate that the assessment designers believed that decoding and

receiving content via the English language were part of the targeted proficiency.[83] With regard to mathematics, a complete mapping from the NAEP mathematics concept of mathematical complexity to the ECD validity argument framework was not judged to be feasible within the scope of this project; however, we have taken elements of the concept of mathematical complexity and combined them with other elements to map into the ECD framework.

***The Concept of Mathematical Complexity in NAEP Mathematics***

This section focuses on the concept of mathematical complexity as found in the 2005 NAEP mathematics framework (National Assessment Governing Board, 2004). The NAEP mathematics framework includes the following:

> Each item written for the NAEP mathematics assessment reflects two major dimensions: mathematical content area . . . and mathematical complexity . . . .
>
> Each NAEP item assesses an objective that can be associated with a content area of mathematics, such as number or geometry. The item also makes certain demands on student's thinking. These demands constitute the mathematical complexity of the item, which is the second dimension of the mathematics framework. The demands on thinking that an item makes—what it asks the student to recall, understand, reason about, and do—are determined based on the assumption that the student is familiar with the mathematics of the task. If a student has not studied these mathematics, the task is likely to make different and heavier demands, and the student may well not be successful on it. .
>
> The complexity dimension is both similar to and different from the levels of mathematical ability (conceptual understanding, procedural knowledge, and problem solving) that were used in the NAEP Mathematics Framework for the 1996 and 2000 assessments. . . . Level of complexity is different from mathematical ability, however, in that complexity describes the mathematical expectations of an item, whereas mathematical ability—along with the associated construct of mathematical power—required an inference about skill, knowledge, and background of the students taking the item. . . .
>
> Moreover, the mathematical complexity of an item is constant; it does not vary depending on the score given for a certain kind or level of response. (prepublication version of the

2005 NAEP mathematics framework, [National Assessment Governing Board, 2001], pp. 18–19)

It appears that the concept of mathematical complexity is intended to stand parallel to, but different from the concept of mathematical ability, the key difference being that mathematical ability is a characteristic of students and mathematical complexity is more nearly a characteristic of the tasks (items). Following are two of the possible mappings of mathematical complexity to the ECD validity argument framework. Both are variations on the same theme of characterizing NAEP mathematical complexity as involving one or more task model variables, which would induce demands for levels of students' knowledge and abilities to solve them.[84]

Option 1. Mathematical complexity as a set of one-or-more task features that generate demand for a focal KSA (e.g., reasoning driver or kind of item) or for the targeted proficiency itself. In this option, mathematical complexity would be virtually synonymous with the term *mathematics driver* (a set of task characteristics that drive demand for a mathematics focal KSA). NAEP language supportive of this option was: "Moreover, *the mathematical complexity of an item is constant*; it does not vary depending on the score given for a certain kind or level of response." (National Assessment Governing Board, 2001, p. 19, emphasis added). Thus, the mathematical complexity of an item would be determined or computed based on some specific set of task features of the item and constitutes a driver of the demand for mathematical ability.

Option 2. Mathematical complexity as the set of (cognitive) demands that are generated by one-or-more task features. NAEP language supportive of this option was: "Each NAEP item assesses an objective that can be associated with a content area of mathematics, such as number or geometry. The item also makes certain *demands on student's thinking*. These *demands constitute the mathematical complexity of the item*, which is the second dimension of the mathematics framework" (p. 18, emphasis added).[85]

It should be emphasized that ECD makes finer distinctions than is commonly done in this portion of an assessment argument, so it is not surprising that a NAEP framework document would not make the distinctions among characteristics of students, tasks, and responses as cleanly or distinctly. If we assume, for a moment, that Option 1 is correct, that mathematical complexity is essentially a mathematical driver, then this would tend to lead to a situation in which the targeted proficiency had only one focal KSA—mathematical ability. That is, if a good mathematical ability KSA is all that it takes to have good mathematics proficiency, then all other

KSAs are nonfocal rather than focal. While for simple analyses of accommodations (e.g., large font and low vision for reading), having only the single focal KSA within the targeted proficiency may be appropriate, this leaves something to be desired when analyzing situations that are more difficult. Specifically, we have found it useful to have at least two focal KSAs, where one KSA is more dominant or central to the targeted proficiency than the others.[86] Instead of using mathematical complexity directly in our mapping to the ECD validity argument framework, we have pursued a different approach.

***Focal KSAs***

Reading and mathematics would each be modeled being constituted of a dominant focal KSA, supplemented by a less-prominent focal KSA. For reading, the dominant focal KSA would be comprehend and for mathematics the dominant focal KSA would be reason. The focus on comprehension for reading is consistent with models for reading comprehension presented earlier in this report. As for mathematics, the NAEP mathematics framework specifically mentions reasoning, once in the general description of the cognitive demands of the assessment (recall, understand, *reason* about, and do, National Assessment Governing Board, 2001, p. 18, emphasis added) and again in the descriptions of moderate complexity items (p. 20) and high complexity items (p. 21).[87]

Let us now consider the less-prominent focal KSAs. For reading the additional focal KSA would be decoding, on the assumption that NAEP reading was intended to include decoding (see earlier discussion of the definition of reading in connection with the example for Tim and being blind). Thus, for this modeling of NAEP reading, we adopted a definition essentially like that of Definition A of reading comprehension that was discussed at length earlier in this report.

For mathematics, the additional focal KSA would be knowledge of the content vocabulary (i.e., mathematics vocabulary). Possible examples for mathematics vocabulary—found in sample items in the NAEP mathematics framework—would include cube (p. 61), vertex (p. 61), multiplying (p. 61), radius (p. 62), cylinder (p. 62), speed (mph) (p. 68), mean (average) (p. 76), perpendicular (p. 80), parallelogram (p. 82), and diagonals (p. 82). One could argue that focusing on vocabulary would be like focusing on factual recall, a skill that NAEP has said is not the major emphasis of NAEP items (All NAEP questions emphasize critical thinking skills and *reasoning* rather than factual recall, 2003 NAEP reading framework, p. 20, emphasis added). Yet recall (possibly meaning, among other things, recall of vocabulary) is mentioned as part of what

NAEP mathematics should measure (2005 NAEP mathematics framework, p. 18) and recall or recognize a fact, term, or property is mentioned first among demands for low complexity items.

*Definitions of reading and mathematics.* Table 20 provides the basic definition of reading and mathematics that we will be working with in this section.[88]

**Table 20**

***Basic Definitions of Reading and Mathematics Targeted Proficiency***

| KSA | Reading | Mathematics |
|---|---|---|
| Comprehend | Good | n/a |
| Reason | n/a | Good |
| Decode | Okay | n/a |
| Know content vocabulary | n/a | Okay |
| Know noncontent vocabulary | n/a | n/a |

*Note.* KSA = knowledge, skills, and abilities.

The KSA levels (from among good, okay, and poor) shown in Table 20 are minimums required in order to have good targeted proficiency. For example, in order to have good reading proficiency, a student would need comprehension at the good level and decoding at least at the okay level.[89]

*Some key features of the model.* The Bayes net model discussed in this section builds on the work discussed earlier in the report. Following is a summary of notable features, with emphasis on changes from the previous model.

1. Allows two major definitions of the targeted proficiency. These are reading and mathematics. The definition of proficiency node is used to switch between these definitions.[90]

2. Includes additional person characteristics (KSAs). In addition to comprehend, reason, decode, know braille codes, feel, see, hear, the model also has know content vocabulary, and know noncontent vocabulary. These additions are consistent with need to be able to model the demands for vocabulary knowledge that is specific to a domain being assessed (content vocabulary) versus not specific (noncontent vocabulary). In addition to providing a richer picture of linguistic and reading-related demands, this distinction is particularly important in modeling the impact of different kinds of dictionary. (The dictionary will be discussed later at some length.)

3. Makes greater use of effect-oriented names for task model variables. That is, some task model variables are named by their effect, particularly of generating demand for some specific KSA. The model uses the term comprehension-reasoning driver instead of kind of item to name the task feature that generates demand for comprehension or reasoning ability. The NAEP mathematical complexity coding might play such a role. The model also uses the term decoding driver instead of frequency of words with multiple contiguous consonants. The variables font size and presentation mode retain their same names and values as in earlier models. While this naming convention has the disadvantage of not being as descriptive of the specific task features, it seems more useful in the present context for communicating the purpose of these variables.

4. Provides additional task model variables. The additional nodes are content vocabulary driver, noncontent vocabulary driver, and dictionary. Content vocabulary refers to words from the domain being assessed. Noncontent vocabulary refers to other words. Both these vocabulary terms have high, medium, and low levels, with high values being generated by features such as the rarity (low frequency of occurrence) of the vocabulary. Dictionary has three values—none, regular, and custom. A regular dictionary contains entries for both content vocabulary and noncontent vocabulary. A custom dictionary contains has been carefully inspected by experts to ensure that it provides useful information about noncontent vocabulary but not about content vocabulary.[91]

5. Provides a node to switch between three specific task model variants, one for reading and two for mathematics. This node provides a convenient way to set some task model variables.

6. Introduces efficiencies to reduce the number of nodes. Because of the structural similarity in the roles filled by the dominant focal KSAs of comprehension and reasoning in reading and mathematics, respectively, we used the same Bayes net nodes for variables of both targeted proficiencies. Specifically, we used three variables called comprehend-reason, comprehension-reasoning driver, and meet comprehension-reasoning demand for both comprehension and reasoning. Because

we don't run analyses for reading and mathematics at the same time, this efficiency strategy does not hinder use of the models.

### *Focal and Nonfocal Knowledge, Skills, and Abilities for Reading and Mathematics*

This section provides a bit more detail on the focal and nonfocal KSAs for reading and mathematics in the context of the model. Let use begin by examining the targeted proficiencies for both reading and mathematics with respect to five cognitive skills:

1. Comprehend: We define this as having two levels—good and poor. We use this in analyses of reading.
2. Reason: We define this as having two levels—good and poor. We use this in analyses of mathematics.
3. Know content vocabulary: This refers to the knowledge of the meanings of words that are relevant to the content or domain being assessed (e.g., mathematics). We define this as having three levels—good, okay, and poor.
4. Decode: This refers basically to the ability to form words from characters. We define this as having three levels—good, okay, and poor.
5. Know noncontent vocabulary: This refers to the knowledge of the meanings of words that are not specifically relevant to the content or domain being assessed. We define this as having three levels—good, okay, and poor.

Note that having only two or three levels of each student model variable is a simplification used in this presentation to simplify the exposition. The 2- and 3-levels approach has the advantages of simplicity of model creation and simplicity of interpretation. However, in an another model—not in the scope of this report—one might use continuous variables, for example, the proficiency variables in an IRT model, to represent some or all of a students' proficiencies.

Let us first consider reading. Table 20 shows that in order for a person to be considered as having good reading ability under this conjectured definition of that construct, they must have both good reasoning skill and okay (or better) decoding skill. Any level of skill—even the lowest level (poor)—is sufficient for the other two skills (know content vocabulary and know noncontent vocabulary). Utilizing earlier terms and conventions, we would says that for reading,

reason and decode are focal KSAs, because they are significant constituents of the targeted proficiency while the other two are nonfocal KSAs.[92]

Let us now consider mathematics. The table shows that in order for a person to be considered as having good mathematical ability, that person must have both good reasoning skill and okay (or better) knowledge of content (mathematics) vocabulary. Any level of skill—even the lowest level (poor)—is sufficient for the other two skills (decode and know noncontent vocabulary). We would says that for mathematics, reason and know content vocabulary are focal KSAs, because they are significant constituents of the targeted proficiency while the other two are nonfocal KSAs.

Having identified the key focal KSAs and a few of the nonfocal KSAs for reading and mathematics, let us consider a broader set of nonfocal KSAs. (Recall that the set of nonfocal KSAs is simply the set of all KSAs that are not focal KSAs.) As we have discussed, what is a nonfocal KSA for one targeted proficiency may be a focal KSA for another. For example, in the previous table, we can see that decoding is a focal KSA for reading but a nonfocal KSA for mathematics. The full set of focal and nonfocal KSAs for this Bayes net are in Table 21.

**Table 21**

***Focal and Nonfocal Knowledge, Skills, and Abilities***

| KSA and levels | Reading | Mathematics |
|---|---|---|
| Comprehend (good, poor) | Focal KSA | Nonfocal KSA |
| Reason (good, poor) | Nonfocal KSA | Focal KSA |
| Decode (good, okay, poor) | Focal KSA | Nonfocal KSA |
| Know content vocabulary (good, okay, poor) | Nonfocal KSA | Focal KSA |
| Know noncontent vocabulary (good, okay, poor) | Nonfocal KSA | Nonfocal KSA |
| Know braille codes (yes, no) | Nonfocal KSA | Nonfocal KSA |
| Feel (yes, no) | Nonfocal KSA | Nonfocal KSA |
| See (yes, partial, no) | Nonfocal KSA | Nonfocal KSA |
| Hear (yes, no) | Nonfocal KSA | Nonfocal KSA |

*Note.* KSA = knowledge, skills, and abilities.

*Task features of the model.* The model's full set of task features with their possible settings is as shown in Table 22. Task features generate demand for cognitive and physical KSAs. Generally, well-crafted task features generate the right amount of focal requirements and

minimize the nonfocal requirements. An accommodation generally involves arranging features of the task performance situation to reduce or eliminate requirements for nonfocal KSAs in which there is a deficit and instead rely on nonfocal KSAs in which there is no deficit.

**Table 22**

***Task Model Variables and Their Levels***

| Task model variable | Levels | Example detail regarding levels |
|---|---|---|
| Presentation mode | Visual text, read-aloud, braille | N/A |
| Comprehension or reasoning driver | High, low | Complexity of sentence structures used in prompts |
| Decoding driver | High, medium, low | Frequency of words with multiple contiguous consonants |
| Font size | Regular, large | Size of font on visual display |
| Content vocabulary driver | High, medium, low | Rarity of vocabulary within content domain (e.g., increasingly specialized mathematics terminology) |
| Noncontent vocabulary driver | High, medium, low | Rarity of vocabulary outside the content domain |
| Dictionary | None, regular, custom | N/A |

*Nonfocal and focal requirements.* Fortunately, during operation of the model, the Bayes net handles the mechanics of determining whether the nonfocal or focal requirements have been met and the basic consequences of those determinations. That is, the implications of all the possible combinations of student and task model variables that we have been discussing have been encoded into the conditional probability matrices of the Bayes net, and can now be evaluated automatically. (Recall that the conditional probabilities are the way of embedding warrants—or at least the machinery for them—into the Bayes net and represent generalizations based on knowledge and experience.) Assuming that the model is correct, the user provides inputs in the form of assessment characteristics and person characteristics and then the machinery of the Bayes net automatically produces the appropriate results.

In order for the machinery to yield the proper results, much knowledge needs to be built into the Bayes net. Knowledge about whether a KSA is a focal KSA or a nonfocal KSA is important, but it is often not enough. The model should have encoded within it. For example, knowledge about the specific levels that are required to perform well (i.e., to have good effective

proficiency) and knowledge about the specific consequences of failure to meet those requirements. These are some of the details that can make model construction both challenging and interesting.

It is important to keep in mind that whether a KSA is nonfocal or focal depends on the definition of the targeted proficiency. On the other hand, whether a KSA is required or not depends on whether the KSA is necessary for good performance in an operational assessment situation.[93] For NAEP reading and mathematics assessments administered under default conditions, examples of nonfocal requirements (i.e., KSAs that are both nonfocal and required) would include the ability to see (in order to access printed text of the assessment) and hear (to be able to hear spoken directions), and so on. As we consider important accommodations, we know that different accommodations cause different nonfocal requirements. For example, administering an assessment via braille, nonfocal requirements would include the ability to feel (tactile sense) as well as knowledge of braille codes.

*Validity for nondisabled individuals under default conditions.* In the basic analysis we assume that nondisabled individuals being assessed under default conditions are being validly assessed.[94] For example, we assume that the assessment content is being displayed as visual text in regular font and that demands for nonfocal skills (be they for decoding, knowledge of both content vocabulary and noncontent vocabulary, and so on. depending on the content area) can be met by the nondisabled individual. Later in this report (in connection with the dictionary situation) we will consider a case in which that assumption is not met.

*A picture of the Bayes net.* Figure 24 shows a graphical representation of the Bayes net. This model, like several earlier ones, corresponds reasonably well to the general schema presented earlier. *Basic findings and interpretation.* Following are some basic characteristics, for the current model involving reading and mathematics. The Bayes model developed for reading and mathematics (drawing upon the NAEP framework) consists of 34 nodes (variables) and handles five accommodations (large font, read-aloud, braille, and two kinds of dictionary). Six accommodation packages (each consisting of two accommodations) are available. The model allows one to specify hundreds of different examinee profiles, including those involving five specific disabilities—low vision, blind, deaf, dyslexia, neuropathy—or combinations of disabilities (e.g., deaf-blind-dyslexia-neuropathy). The model allows the user to specify two different definitions of the construct (reading, mathematics) and hundreds of different task

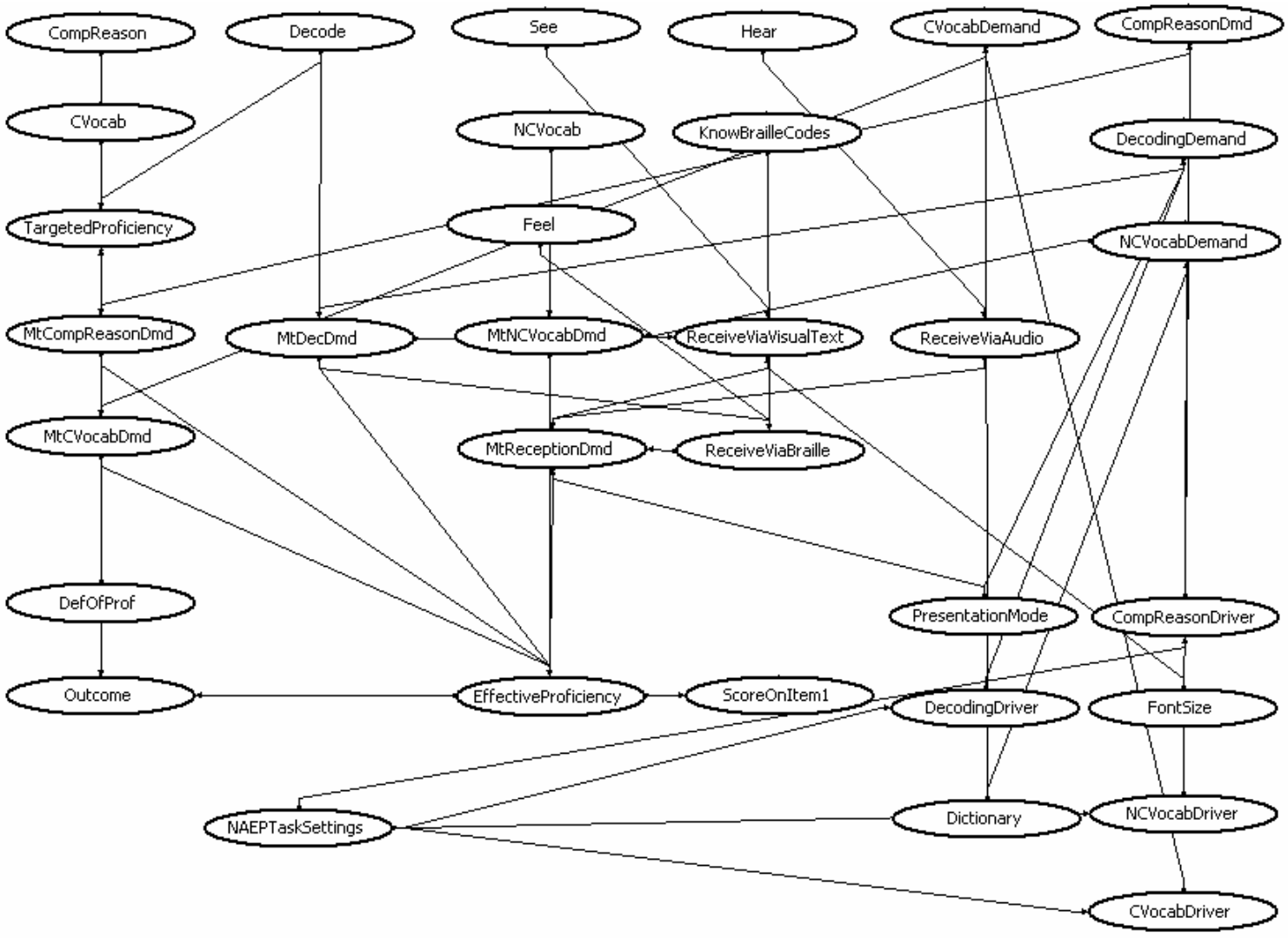


***Figure 24.*** **A picture of the Bayes net.**

performance settings (sets of testing conditions). Counting the variations in examinee profiles (including disability), task performance situations (including accommodations), and definitions of the construct, the model allows one to analyze well over 100,000 situations in terms of the likely validity of the scores that would result.[95]

Table 23 compares the current model to two earlier models, one for Sue and low vision and the other for Tim and being blind.

**Table 23**

***Comparison Between Bayes Net Models***

| | | Bayes net model | | |
|---|---|---|---|---|
| Num | Description | 1. RC: Sue and low vision | 2. RC: Tim and being blind | 3. Reading and mathematics |
| 1 | Number of accommodations | 1 (large font) | 3 (large font, read-aloud, braille) | 5 (large font, read-aloud, braille, regular dictionary, custom dictionary) |
| 2 | Packages (each consisting of two or more accommodations.) | 0 | 0 | 6 |
| 3 | Single disabilities for which the model supports the validity of the accommodation (e.g., low vision, blind, deaf, dyslexia, neuropathy) | 1 (low vision) | 5 | 5 |
| 4 | Analyzes accommodation packages with individuals with multiple disabilities (e.g., deaf-blind, blind-neuropathy, blind-dyslexia, low-vision-deaf) | No | Yes | Yes |
| 5 | Definitions of the targeted proficiency | 1 (RC) | 2 (RC-A, RC-B) | 2 (reading, mathematics) |
| 6 | Number of nodes in Bayes net | 9 | 23 | 34 |
| 7 | Number of unique situations that can be analyzed | 24 | 10,000+ | 100,000+ |

*Note.* RC = reading comprehension.

Table 24 displays the basic findings of running the reading and mathematics models for seven cases involving diverse disability statuses (nondisabled, blind, low vision, dyslexia[96]) and either reading or mathematics. Three accommodations (large font, read-aloud, braille) are involved. Generally we see that NAEP practice is confirmed (last column), as evidenced by the fact that, for individuals with disabilities, NAEP allows accommodations that are valid (according to the model) and disallows ones that are not valid. The word consistent in the last column of the table indicates that there appears to be a general consistency between NAEP accommodation policy and the model-based estimate of validity (generated via the Bayes net). In order to understand the significance of these consistencies, we need to think more about the current practices for providing accommodations as well as the limitations of the modeling approach. Except where an accommodation is forbidden, NAEP appears to try to fulfill the accommodation requirements of the IEP team. For example, NAEP has a list of frequently provided accommodations and may have a list (not available in the framework documents) of other accommodations that may be provided on an infrequent basis. Very few accommodations are actually restricted from being used (e.g., read-aloud on the reading assessment is not allowed). Because NAEP depends heavily on the IEP team, it is possible that NAEP may allow an accommodation that would allow a student to validly participate, but the IEP team may not require it. For example, the expense of the accommodation may have inhibited specification of the accommodation in the IEP or prior use of the accommodation in prior classroom situations. It is not clear whether IEP teams would be more effective in crafting IEPs that were more conducive to inclusion if they had access (well ahead of time) to a more complete list of possible NAEP accommodations. Note that Cases 4 and 11 are cited as not necessarily having consistency between NAEP practice and the model-based results. Both are situations in which an individual with dyslexia and another disability (low vision and blind) are attempting to access reading assessment content in a way that requires decoding (large font visually displayed text in one case and braille and in other). In both cases, the model says that the results are invalid. This is because even though both individuals have true outcomes (true-negative), the model considers the assessment conditions inappropriate since they reception demand was not met (due to dyslexia). Obviously, these situations do not necessarily signify a true inconsistency between the model results and NAEP practice, since, NAEP would rely on the IEP committee's recommendations, which may or may not recommend that the student be included with these accommodations.

**Table 24**

***Basic Findings***

| Case # | Accommodation | Person profile (summary) [a] | Definition of targeted proficiency | Model-based validity [b] | NAEP allows accommodation if called for by IEP | Comment on consistency between NAEP policy on allowed accommodations and model-based validity status [c] |
|---|---|---|---|---|---|---|
| 1 | None | Nondisabled | Reading | Valid | No | Consistent. No accommodation necessary |
| 2 | None | Nondisabled | Mathematics | Valid | No | Consistent. No accommodation necessary |
| 3 | Large font | Low vision, with okay decoding | Reading | Valid | Yes | Consistent |
| 4 | Large font | Low vision, dyslexia | Reading | Invalid | Yes | Perhaps not consistent. Model assumes that large font, by itself, would not ensure that reception demand is met, due to dyslexia. |
| 5 | Large font | Low vision, dyslexia | Mathematics | Valid | Yes | Consistent. Due to low decoding requirement of mathematics, poor decoding does not prevent reception demand from being met. |
| 6 | Read-aloud | Blind, dyslexia | Reading | Invalid | No | Consistent |
| 7 | Read-aloud | Blind, dyslexia | Mathematics | Valid | Yes | Consistent |
| 8 | Read-aloud | Dyslexia | Reading | Invalid | No | Consistent |

*(Table continues)*

Table 24 (continued)

| Case # | Accommodation | Person profile (summary) [a] | Definition of targeted proficiency | Model-based validity [b] | NAEP allows accommodation if called for by IEP | Comment on consistency between NAEP policy on allowed accommodations and model-based validity status [c] |
|---|---|---|---|---|---|---|
| 9 | Read-aloud | Dyslexia | Mathematics | Valid | Yes | Consistent |
| 10 | Braille | Blind, okay or good decoding | Reading | Valid | Infrequently if at all | Consistent, if offered |
| 11 | Braille | Blind, dyslexia | Reading | Invalid | Infrequently if at all | Perhaps not consistent. Model assumes that braille, by itself, would not ensure that reception demand is met, due to dyslexia. |
| 12 | Braille | Blind, dyslexia | Mathematics | Valid | Infrequently if at all | Consistent, if offered. Due to low decoding requirement of mathematics, poor decoding does not prevent reception demand from being met. |

[a] The person profile assumes that there is no other disabling condition than that listed and that all other KSAs are favorable to validity. [b] The definition of validity is that for individuals with both good and poor targeted proficiency, the results of the model show that (a) deductive outcomes are true (true-positive or true-negative), (b) reception demand is met, and (c) appropriate focal requirements are maintained. [c] This analysis makes the simple assumption that poor decoding ability is a defining characteristic of dyslexia. It also assumes that individuals with poor decoding ability have a very high probability of being diagnosed as having dyslexia. These examples also make the simple assumption that nature of dyslexia is fundamentally the same, regardless of any additional disability. A more nuanced analysis might modify these assumptions.

It may not be practical for NAEP to do anything differently than it does now in this matter. Yet this kind analysis can at least point out situations that conflict with specific definitions of validity.

Keeping in mind the desirable goal of including as many individuals with disabilities in NAEP as possible, one wonders if there are rational ways of allowing the read-aloud accommodation, even though it appears to be invalid in some situations. It should be noted that the Bayes net recognizes that where good reading ability is defined as requiring at least okay decoding, then a person with poor decoding ability would receive an unfair advantage by using a read-aloud accommodation. However, what about individuals with okay or good decoding ability? Should the read-aloud accommodation be allowed for individuals that independent evidence suggests already have okay or good decoding ability? For example, imagine a nondisabled student who is an excellent reader when using the printed page. He has good decoding skills. He then becomes blind and relies heavily on the read-aloud accommodation in academic situations. Or imagine a student who has been blind for a lifetime but who can demonstrate good decoding skills in a diagnostic test that has her pronounce (and perhaps define) words that are spelled out auditorily. In order for NAEP to do its job in measuring reading achievement, the question becomes: Is it essential that examinees demonstrate their decoding ability during the assessment, or in the interest of inclusion, can independent evidence of decoding ability suffice?[97] To allow the use of such an approach would amount to a revision of assumptions underlying one of our validity criteria. One of our criteria was that the accommodation should maintain (for individuals with disabilities) the same focal requirements experienced by nondisabled examinees who are validly assessed under default conditions; our implicit assumption was that these focal requirements upon the nondisabled user were experienced during the assessment, not before the assessment. Allowing other diagnostic information about the examinee's ability to meet a focal requirement would seem to violate that assumption. Use of such information is consistent with the practice of considering each individual with a disability on a case-by-case basis. Yet a variety of issues would arise with such an approach, among them the reliability of the external measure and the difficulty of certifying such diagnoses. While the ECD approach cannot, of itself provide an answer to the feasibility of such an approach, it can provide a framework for asking the questions and examining the implications of different answers. Providing an answer to such a question would require an

examination of how such information could be acquired, whether it could be done at a reasonable cost, and whether such a policy would be consistent with the uses of NAEP result data.

The foregoing findings suggest that the model is capable of representing a wide array of accommodation-related situations and that its results generally correspond well to NAEP practice. In interpreting these results we need to keep in mind that many simplifications and assumptions were made in the construction of the model and, therefore, that there are limitations in the inferences that one can draw either about the quality of the ECD validity argument approach or NAEP accommodation practices.[98] Realistically, the results of running these models should be viewed as hypotheses that one would attempt to confirm or disconfirm with other evidence.

*Relationship between the modeling activity and empirical research.* It is useful at this point to discuss the important roles that empirical research plays in the creation and use of ECD validity arguments. Especially in our modern educational environment that challenges us to produce research-based findings, it is useful to consider examples of ways in which the kind of work described in this report connects with research, especially empirical (including quantitative) research. This section discusses ways in which empirical research can inform the creation, use, and refinement of models.[99] Let us consider a few examples.

First of all, a well-constructed model embodies research theories and findings. The structures (e.g., schema and other knowledge representations) seek to exemplify the thinking–including research-based thinking—of experts and researchers in evidentiary reasoning, educational measurement, accessibility, and subject-matter domains. For example, the conditional probabilities entered into the Bayes net should be based on theory and experience, including empirical research. ECD encourages us to make explicit the linkages between research findings and features of models. Indeed, an ECD-based design management system could include facilities for storing documentation (i.e., backing for the warrants) and relating it to the related assessment design structures (Frase et al., 2003).

Another example of research playing a role in model creation is the fact that an appropriately defined targeted proficiency will include KSAs that our research has found to exist in individuals who are successful in criterion situations (e.g., situations to which the assessment scores are likely to be used to relate to success variables). The assessment planner would thus be

a consumer of research knowledge about the skills and abilities used by individuals during successful performance in criterion situations.[100]

Empirical research can guide the refinement of a model. Suppose the model indicates that use of a certain type of accommodation is invalid for use with individuals with a certain person profile (e.g., a specific set of nonfocal/focal KSAs,). Yet suppose that empirical research shows that individuals with that profile who use the accommodation have scores that correlate just as highly with an external criterion as do scores from nondisabled individuals taking the assessment under default conditions and whose scores, therefore, are believed to be valid. Such research would invite consideration of the cause or causes of this discrepancy, and may involve refinement or correction of the model.[101]

Consider another example of how empirical research can help refine models. The models we have discussed indicate that a person with unimpaired vision will perform equally well with regular font or large font. They also indicate that a person with low vision and having the same level in the targeted proficiency as the nondisabled person would perform equally well using large font. (These model-based findings seem to suggest that large font might be offered to nondisabled individuals if they desire it.) 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 use of the model, it may or may not be important to attempt to model such subtleties. Even if such fine points of research knowledge do not become explicit parts of model creation, they can play a role in guiding the wise use of the results produced by the model.[102]

Another way that empirical research can help refine models is by helping identify nonfocal or focal requirements that are excessive. An example of this kind of issue concerns excessive linguistic load in items.

*Linguistic load of mathematics assessments.* The issue of excessive linguistic load on nonreading assessments, such as mathematics, can be a significant issue for individuals with disabilities as well as English-language learners.[103] It may also be an issue for individuals who are not classified as having disabilities who have some difficulty in reading. Abedi (2002) has stated that “Reducing unnecessary language complexity of test items helps ELL students (and to

some extent SDs [student with disabilities]) to present a more valid picture of their content knowledge."

In a study of 1992 NAEP mathematics and science assessments, Abedi and his colleagues grouped test items into linguistically complex and less-complex items. They found that ELL students had higher scores on the linguistically less-complex items (Abedi, Lord, & Plummer, 1995, cited in Abedi, 2002). Consistent with that finding is research on mathematics items with relatively low language demand, such as mathematics computation. With such items, the performance gap between ELL students and native speakers of English decreases or even disappears (Abedi, Leon, & Mirocha, 2001, cited in Abedi, 2001, p. 2).

In other studies, Abedi and colleagues have examined the impact of linguistic modification, which among other things involves using familiar (as opposed to unfamiliar or infrequently used) noncontent vocabulary (e.g., video game instead of census; Mack's company instead of a certain reference file). Studies have found that ELL students preferred linguistically modified NAEP mathematics items over the original items and performed better on them than the unmodified ones (Abedi, Lord, & Plummer, 1997, cited in Abedi, 2003). His studies have found that reducing language complexity helps narrow the performance gap between native English speakers and ELL students (Abedi, 2001, p. 3), thus improving the performance of ELLs but with little or no improvement in the performance of native English speakers. Abedi recommends that assessment developers "reduce linguistic complexity during development and improvement of all large-scale assessment programs" (Abedi, 2001, p. 3).

While reduction in linguistic complexity is a step involving a change in the default or general characteristics of the task performance situation, Abedi also recommends an accommodation (i.e., a deviation from the default conditions) as another way of accomplishing a similar end. Specifically, based on research in assessments of domains like mathematics and science (but not reading), he suggests consideration of the use of a custom dictionary that includes only noncontent vocabulary.[104] He notes that a regular dictionary that includes both content words and noncontent words "may provide ELL students an unfair advantage on certain types of tests" (Abedi, 2001, p. 3).

The NAEP mathematics framework appears to reflect an understanding of the importance of keeping unnecessary linguistic complexity to a minimum, while avoiding alteration of the construct being measured. This awareness of the importance of avoiding excessive linguistic

demands usually appears under the term plain language. The NAEP mathematics framework describes the use of plain language editing procedures (National Assessment Governing Board, 2001, p. 7; see also p. 15, 22–23) as way to bring this about:

> Plain language is a writing and editing tool designed to clearly convey meaning without altering what items are intended to measure. All items should use plain language. Even when the intent of the item is for students to define, recognize, or use mathematics vocabulary correctly, the surrounding text should be in plain language. Plain language guidelines often increase access and minimize confusion for students. * * * Use high frequency words as much as possible. (pp. 22–23)

Thus, the goal of plain language editing is essentially to minimize unnecessary complexity without changing what the assessment is intended to measure.

*An example of excessive linguistic load.* The NAEP mathematics framework displays a good awareness of the issues of linguistic load. Several pages describe guidelines for avoiding unnecessary linguistic complexity. But the framework provides little to no detail as to whether efforts to limit unnecessary linguistic complexity have been successful. The possibility then arises that the actual requirement for linguistic ability due to linguistic complexity is greater than intended (i.e., greater than what is necessary). With this possibility in mind, let us now take an example of one kind of excessive linguistic load and discuss how that might be dealt with.

Consider Table 25. Column 1 shows the four KSAs and column 2 shows the minimum values defining the mathematics targeted proficiency. Recall that KSAs with values of okay and good—reason and know content vocabulary—are considered focal KSAs. The other two—decode and know noncontent vocabulary—are nonfocal KSAs because even the lowest level of KSA (poor) is sufficient. Column 3 shows the (maximum) demands that can be satisfied by the levels defined in column 2. The goodness of the match can be recognized from our convention that a good level of a KSA satisfied a high demand (or requirement), an okay level satisfies a medium level, and a poor level satisfies a low level. Thus, generally speaking, task demands at the levels shown in column 3 are ideal for eliciting evidence about whether someone has good mathematics proficiency. For example, it is a set of task demands that (assuming all other circumstances are favorable to valid measurement) will tend to yield valid results (e.g., true outcomes [e.g., true-positives and true-negatives] as opposed to false ones). (The fifth row in the table indicates the likely validity).

**Table 25**

***Linguistic Demands in a Mathematics Assessment, Using a Variety of Conditions***

| 1. KSA | 2. Definition of mathematics proficiency | 3. Ideal demand (no accommodation) | 4. Excessive demand for know noncontent vocabulary (no accommodation) | 5. Demand when using custom dictionary accommodation | 6. Demand when using regular dictionary accommodation |
|---|---|---|---|---|---|
| Reason | Good | High | High | High | High |
| Know content vocabulary | Okay | Medium | Medium | Medium | Low |
| Decode | Poor | Low | Low | Low | Low |
| Know noncontent vocabulary | Poor | Low | Medium | Low | Low |
| Valid or invalid | | Valid | Invalid | Valid | Invalid |
| Nature of bias/invalidity (as revealed in the operation of the Bayes net model) | | n/a | Bias against persons for whom know noncontent vocabulary = poor | n/a | Bias in favor of persons for whom know content vocabulary = poor |

Column 4 is like column three except that it shows a specific kind of excessive linguistic load. Specifically, instead of having low demand for noncontent vocabulary, it has medium demand. This situation would be biased against individuals who possess the KSA levels in column 2 (the definition of mathematics proficiency) but who would be unable to demonstrate that proficiency because the excessive requirement (demand) for knowledge of noncontent vocabulary. It seems reasonable to think that many individuals without and with disabilities, including some with various kinds of learning disabilities, would have their scores negatively affected. (Note that this case of excessive linguistic load deviates from our earlier assumption of nondisabled individuals having their abilities validly measured.)

How should such an excessive linguistic load be addressed? Let us consider the methods we discussed earlier. We could rewrite the items to reduce the linguistic load to the proper level (through plain language or linguistic modification). Or we could provide an accommodation of a custom dictionary (i.e., dictionary of noncontent words) that reduces the requirement for knowledge of noncontent vocabulary to the proper level. This situation is shown in column 5.

In column 6 we see what would result from using a regular dictionary (i.e., one that includes both content vocabulary and noncontent vocabulary) as an accommodation. We have modeled this dictionary as keeping the demand for both content vocabulary and noncontent vocabulary at a low level. Such a feature would result in a bias in favor of individuals with poor knowledge of content vocabulary who, though having poor mathematical proficiency (per the definition of mathematical proficiency), are able to perform well on the tasks due to the entries for mathematics content vocabulary that they find in the regular dictionary. This is an example of an accommodation inadvertently reducing demand for focal KSA (know content vocabulary).

*Influence of dictionaries on demand for vocabulary knowledge.* Let us now consider in greater detail some of the reasoning that goes into representing dictionary-related issues in the Bayes net. This is provided as additional detail for those who desire it. Let us begin by examining how the model treats the impact of the dictionary variable on demand for vocabulary knowledge. *Dictionary* has three possible values: (a) none, (b) custom (where a custom dictionary contains only noncontent words), and (c) regular (where a regular dictionary contains both content words and noncontent words).

Table 26 shows the impact of dictionary on demand for knowledge of noncontent vocabulary. The key idea here is that either kind of dictionary (custom or regular) has the

capacity to keep the (nonfocal) requirement for knowledge of noncontent vocabulary to a minimum (i.e., low). (This makes sense because if a person can use the dictionary to know the meaning of noncontent words, then they require less ability in that KSA.) Table 26 shows the conditional probabilities for the node or variable called noncontent vocabulary demand, meaning the demand for knowledge of noncontent vocabulary. This variable is modeled as dependent on (having as parents) the variables dictionary and noncontent vocabulary driver. As discussed earlier, noncontent vocabulary driver is an effect-based name for one or more task features that drive demand for noncontent vocabulary knowledge; an example of such a feature would be the rarity (familiarity) of the noncontent vocabulary. We see that where the value of noncontent vocabulary driver is high, medium, or low (based on features of the test content), the demand for noncontent vocabulary knowledge is correspondingly high, medium, or low when no dictionary is present. However, when there is a dictionary present (either regular or custom), the demand for knowledge of noncontent vocabulary knowledge is low.

**Table 26**

***The Impact of Dictionary on Demand for Knowledge of Noncontent (NC) Vocabulary***

| Parent node(s) | | Noncontent (NC) vocabulary demand | | |
|---|---|---|---|---|
| Noncontent (NC) vocabulary driver | Dictionary | High | Medium | Low |
| High | None | 1.0 | 0.0 | 0.0 |
| | Custom | 0.0 | 0.0 | 1.0 |
| | Regular | 0.0 | 0.0 | 1.0 |
| Medium | None | 0.0 | 1.0 | 0.0 |
| | Custom | 0.0 | 0.0 | 1.0 |
| | Regular | 0.0 | 0.0 | 1.0 |
| Low | None | 0.0 | 0.0 | 1.0 |
| | Custom | 0.0 | 0.0 | 1.0 |
| | Regular | 0.0 | 0.0 | 1.0 |

Now let us examine the impact of dictionary on knowledge of content vocabulary. Table 27 shows the conditional probabilities for the node or variable called content vocabulary demand meaning the demand for knowledge of content vocabulary. This variable is modeled as depending on (having as parents) the variables dictionary and content vocabulary driver. We see that where the value of content vocabulary driver is high, medium, or low, the demand for noncontent vocabulary knowledge is correspondingly high, medium, or low when no dictionary is present. However, when there is a regular dictionary present, then the demand for knowledge

of content vocabulary is low. However, whereas a regular dictionary reduces demand for knowledge of content vocabulary, the custom dictionary does not impact it at all (i.e., it has the same impact as no dictionary at all [dictionary = none]).

Thus, the custom dictionary has a very specific impact on demand for knowledge of noncontent vocabulary and does not affect demand for knowledge of content vocabulary. It is this selective impact that makes a custom dictionary a valuable approach for reducing linguistic load without affecting demand for content vocabulary (construct-relevant vocabulary).

**Table 27**

***The Impact of Dictionary on Demand for Knowledge of Content (C) Vocabulary***

| Parent node(s) | Dictionary | Content (C) vocabulary demand | | |
|---|---|---|---|---|
| Content (C) vocabulary driver | | High | Medium | Low |
| High | None | 1.0 | 0.0 | 0.0 |
| | Custom | 1.0 | 0.0 | 0.0 |
| | Regular | 0.0 | 0.0 | 1.0 |
| Medium | None | 0.0 | 1.0 | 0.0 |
| | Custom | 0.0 | 1.0 | 0.0 |
| | Regular | 0.0 | 0.0 | 1.0 |
| Low | None | 0.0 | 0.0 | 1.0 |
| | Custom | 0.0 | 0.0 | 1.0 |
| | Regular | 0.0 | 0.0 | 1.0 |

*Summary regarding linguistic load in NAEP mathematics.* Following are key points about linguistic load in NAEP mathematics:

1. Excessive linguistic load seems to have been an issue historically for NAEP mathematics.
2. The NAEP mathematics framework shows significant awareness of the issue (e.g., the requirement for plain language) but provides little or no detail about the degree of success in addressing the issue. Whether this issue has been adequately addressed could be a matter for research along the lines of experiments by Abedi and his colleagues.
3. If excessive linguistic load is still an issue after efforts to implement plain language practices, then the provision of a custom dictionary (containing noncontent words) might be considered for implementation.

4. If a custom dictionary is to be implemented, consideration should be given to the relevant audiences (with/without disabilities; ELLs/native speakers of English). Depending on the needs and feasibility, the feature might be implemented either as an accommodation for eligible students or part of a new set of default features (e.g., as a universal design feature).

5. The ECD validity argument approach seems capable of representing (a) accommodation-related issues as well as (b) issues related to the possibility of changes to default administration conditions (e.g., universal design features). It also shows potential in dealing with issues affecting English-language learners.

## Discussion and Conclusions

This section briefly summarizes some of the key cases that we have examined. This summary relies heavily on the basic distinction between focal and nonfocal KSAs. We began with a case in which the individual performed poorly despite good ability (Sue and low vision). Poor performance resulted from a deficit in a nonfocal KSA (sight) due to default testing conditions (font size = regular). The accommodation of large font reduced the nonfocal requirement in which there was a deficit (sight) without changing the focal requirement. Thus it was judged to be a valid accommodation in that situation. Later analysis of Sue and low vision suggested that large font might also be an acceptable universal design feature that could be offered to virtually anyone.

We briefly discussed a situation involving Carl in which there seemed to be a fairly obvious conflict between the targeted proficiency and the accommodation (spell checker for a spelling test). The accommodation would reduce focal requirements for spelling ability, thereby giving rise to false-positive results and, hence, would not be a valid accommodation.

In the case of Tim and being blind involving reading comprehension, we began with an assumption that decoding was a focal KSA (Definition A of reading comprehension). A read-aloud accommodation would eliminate the nonfocal requirement for sight (in which there was a deficit) and increase the reliance on a nonfocal KSA in which there was no deficit (hearing). However, this accommodation would inadvertently reduce the focal requirement for decoding, thereby giving rise to potential false-positive results and, hence, supporting the idea that read-aloud would not be a valid accommodation.

In the course of that discussion, it was said that using a braille accommodation might be a valid accommodation because it overcomes the accessibility barrier (for an examinee who reads braille) without changing the focal requirement for decoding ability.

We saw that if one changes the definition of the targeted proficiency by including only comprehension but not decoding (changing from Definition A to Definition B) then the read-aloud accommodation results in a true outcome, which would be evidence in favor of read-aloud being a valid accommodation. However, such an accommodation could raise fairness and comparability issues for individuals without disabilities who do not receive the accommodation yet have trouble satisfying the nonfocal requirements for decoding. (This would be an issue where some examinees categorized as nondisabled have poor decoding ability.)[105] If there were such fairness and comparability issues, then this would appear to violate our assumption that nondisabled individual receiving the assessment in default conditions is validly assessed. Such considerations introduce uncertainty into the assertion that the accommodation would be valid.

In the area of mathematics, we examined situations in which mathematics proficiency was defined as requiring both good reasoning and okay-or-better knowledge of mathematics vocabulary. More specifically, we examined what would happen if the requirement for knowledge of nonmathematics vocabulary were too high. It was noted that a regular dictionary having both mathematics and nonmathematics vocabulary might provide an unfair advantage to the recipient by reducing the requirement for knowledge of content (mathematics) vocabulary. On the other hand, a custom dictionary having only nonmathematics words would not present this problem. Thus, a custom dictionary reduced the nonfocal requirement for knowledge of nonmathematics words without reducing the focal requirement for knowledge of mathematics words. However, if it were the case that the requirement for knowledge of nonmathematics vocabulary was excessive for all examinees, the provision of a custom dictionary might be appropriate for all examinees (as a universal design feature). Indeed this example is not readily classified as a conventional accommodation situation. For example, this situation violates our assumption that individuals without disabilities were validly assessed under default conditions, since there was excessive demand for nonmathematics vocabulary. (Furthermore, we did not try to evaluate what low level of knowledge of noncontent vocabulary is indicative of a disability.)

Another possible approach to addressing the problem of excessive requirements for noncontent vocabulary would be through rewriting items with simpler noncontent words. The

dictionary example is important because it begins to move beyond accommodations into a wider set of strategies for addressing nonfocal (and focal) requirements. It helps show the robustness of the validity argument framework for dealing with a wider set of strategies.

### *Summary of the Approach*

This report has basically explained a series of models, beginning with the Toulmin diagram and then building increasingly elaborate Bayes net models that are applicable to NAEP reading and mathematics. Having been through this exercise, we can summarize the approach.

Essentially, the approach seeks to maximize inclusion of individuals with disabilities while maintaining or enhancing validity. The approach involves creating and using a model of the validity argument for an assessment, with emphasis on parts related to testing accommodations. Based on inputs such as characteristics of the person (including disabilities) and characteristics of the assessment (the accommodation provided, what the assessment is intended to measure, etc.) the model provides a projection of the validity of the interpretations that would arise from scores for various combinations of student and assessment configurations.

The structures for accomplishing this goal involve a validity argument structure for the assessment. Creating a model is guided by the goal of explicitly and accurately defining the relationship between the claim (the student's level of targeted proficiency) and data (scores). The relationship can be disrupted by causes for alternative explanations (other than one's level in the targeted proficiency) for the occurrence of the data (scores); such explanations represent potential threats to validity. Addressing accessibility issues will typically focus on modeling unfair disadvantages caused by accessibility problems (e.g., reception demand not being met) while being alert to potential for unfair advantages for the examinee inadvertently caused by accommodations being provided. Theory and experience (including empirical research), as well as societal values (such as fairness and validity), should inform (and provide rationales for) the creation and use of the model.

### *Key Steps in the Approach*

This section provides a summary of the approach and a set of suggested steps for using the approach to increase inclusion in an assessment while maintaining validity. Major steps also refer to the portion of the ECD argument that is most involved in that step. These steps can often be done in a somewhat different order and are often done in iterations:

1. Identify possible areas for improvement of inclusion and accessibility.

Give special consideration to providing accommodations that fulfill existing accessibility standards, for which there is promising research, or that are commonly used in schooling or other criterion environments.[106] Identify additional purposes, constraints, and resources for assessment.

2. Define what basic indices of validity one will use.

Identify what summary indices of validity—or other indices of assessment quality—one will examine. These could include

- indices related to accessibility (e.g., meet reception demand, meet demand for recording answers, meet demand for working quickly);
- fidelity to intent (e.g., match between the person's targeted proficiency and effective proficiency);
- maintenance of the intended demand for focal KSAs (i.e., focal requirements); and
- adequate empirical relationship between accommodated scores and performance in criterion situations.[107]

3. Construct the model of the validity argument, with emphasis on parts related to disability.

    - Define relevant profiles of individuals of interest. Give consideration to characteristics related to disability (especially reading-related skills) and language status. Among the set of relevant profiles, it is suggested that one include the profile of an individual without any disability as well as individuals with diverse disabilities.
    - Identify the task performance situations, including both the default situation and deviations from it (e.g., accommodations). Focus on features that most directly affect demand for examinee KSAs.
    - Identify the key focal KSAs and nonfocal KSAs. Distinguish between KSAs that are essential parts of the targeted proficiency (focal KSAs) and those that are not (nonfocal KSAs).

- Further define the targeted proficiency. Identify the level of each focal KSA needed for a person to be considered as having the targeted proficiency.[108] The combination (conjunction) of levels for all focal KSAs is essentially a definition of the targeted proficiency.[109]

- Define effective proficiency using a range of situations selected earlier. Define how effective proficiency (performance under actual assessment conditions) is affected by the different combinations of person profiles (sets of characteristics) and task performance situations. Include consideration of individuals without disabilities taking the assessment under default conditions. Use appropriate psycho-physical models to ensure adequate model accuracy and simplicity.

4. Run the model.

Run the model using a set of person characteristics and a set of assessment characteristics as inputs. Obtain outputs in terms of the validities of the results that would likely result. Recall that each time one sets a node (variable) to a particular value, this results in posterior probabilities for all values of all other nodes in the Bayes net. These posterior probabilities (or a selection or summary of them) are the results. They are interpreted as implications for the remaining variables, if the variables that have been set were known to take the values they were set at.

5. Verify the results then refine and redo Steps 1 through 4 if necessary.

Verify that the results are reasonable and that the model is adequate for its intended purpose(s). An important reference point for interpreting the results of runs involving individuals with disabilities is a run of a nondisabled person taking the assessment under default (standard) conditions; finding validity from such a run simplifies the argument over a finding of invalid results. Checking that the results are reasonable is an ECD model verification or validation process, specifically addressing the degree to which the design concords with best practices, standard procedures, and expert guidance.

6. Apply the results.

Among those accommodations that would yield valid results, implement those that are feasible given resource and other constraints. This means using ECD, CAF, and operational assessment system elements to implement the accommodations policy that has been determined.

### *A Larger Context*

In this discussion section, let us pause for a moment and briefly put accommodations into the larger context of the variety of ways that assessment planners can help ensure assessment validity. It is hoped that the ECD techniques will be seen as useful not only in reasoning about accommodations, but also of potential use in thinking through the various complementary approaches for ensuring validity.

The discipline of thinking in terms of focal and nonfocal KSAs is an important tool to help us see accommodations as one part of a multipart solution to the challenges of creating opportunities for learning and growth. Focal KSAs are part of what we want to measure, so during an assessment (or segment of an assessment) we typically do not seek to modify them. Reasonable causes for improvements in focal KSAs might include instruction, training, insight, and maturation. On the other hand, we generally seek to minimize nonfocal requirements of an assessment. Accommodations are a major method for accomplishing this. Let us go beyond the strategy of accommodations and think about other ways to help someone meet the nonfocal requirements of an assessment. Following are several possibility strategies:

1. Increase examinee capability in nonfocal KSAs. For example, familiarization and practice materials as well as legitimate coaching are usually appropriate ways to increase test taker capacity to satisfy the nonfocal requirements of an assessment.

2. Reduce or eliminate nonfocal requirements by altering the task performance situation. This is often done through what are termed accessibility features.

    - Via accommodation. Accommodations typically reduce or eliminate requirements from nonfocal KSAs in which there is a deficit, seeking to replace them with requirements for nonfocal KSAs in which there is no deficit. Accommodations typically require prior approval. Accommodations have been the major focus of this report.[110]
    - Via reduction in the routine nonfocal requirements of the assessment situation. Such changes can make an assessment easier to access for both disabled and nondisabled individuals whose nonfocal KSAs would not otherwise meet the requisite levels. Such changes may fall under the heading of universal design features (Thompson, Johnstone, & Thurlow, 2002).[111]

3. Revise eligibility criteria for taking the test. While one should strive to include all individuals with disabilities, there may be some individuals whose abilities cannot be measured with a given assessment—even using the most state-of-the-art accommodations or universal design features. Alternate assessments may be appropriate for such individuals.

4. Change the definition of the targeted proficiency. This change would apply to all examinees using the assessment. Such an action automatically raises issues of comparability between results obtained using the old versus the new definition so cannot be undertaken lightly. But by changing the definition of the targeted proficiency, one may change the set of nonfocal requirements, thereby making it easier for individuals with disabilities to demonstrate their targeted proficiency. Obviously, changes to the definition of a targeted proficiency will be easiest to make where the assessment is still in the design stage. Permanent changes in the definition of the targeted proficiency should not be undertaken without due consideration of the impacts on the use of assessment results.[112]

An effective validity framework should help us sort through these options. While our focus of this report has been on assessment accommodations for individuals with disabilities, we should not ignore the other aforementioned strategies for dealing with deficits in nonfocal KSAs. Further, we should try to see these strategies as related in principled ways to our strategies for increasing learners' levels in focal skills.[113] It is beyond the scope of this report to explicate how to use ECD to evaluate in detail the different strategies for satisfying nonfocal requirements or for ensuring appropriate focal requirements for assessments or instruction for different purposes. However, we believe, that by making the validity argument more explicit, we can begin to address this larger set of issues in more principled and efficient ways.

### *Conclusion*

This report has described the use of evidence-centered assessment design (ECD) for accommodations for individuals with disabilities in NAEP reading and mathematics. We have focused on ways of thinking, structures for representing knowledge, and tools for supporting reasoning about assessment accommodations. We believe that application of ECD can help

NAEP reconcile the need for accessible assessments with the need to maintain or strengthen the validity argument for interpretations arising from assessment scores.

The argument for an assessment might be summarized as including (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 score levels (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 has been compromised (Messick, 1989). An example of an alternative explanation for poor performance by an individual with a disability would be 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., blind). 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.

We have focused on building argument structures that might help anticipate and address these alternative explanations, particularly as they relate to test takers with disabilities. Using a Bayes net tool, we have developed runnable models of these arguments, including structures for representing a significant range of alternative explanations, based on simple components. While such models cannot automatically make key decisions, they explicate the underlying assessment argument that is to be embedded in the more obvious elements of an assessment. We used this form of representation for a variety of cases, culminating in a Bayes net model for NAEP reading and mathematics. While such Bayes nets do not capture every nuance or every consideration in making an accommodation, they do capture some of the most important ones.

The approach described in this report is not intended to automate critical accommodation decisions but may make help assessment planners and others involved with accommodation policies to think about and make such decisions. While key elements of the approach can be carried out without use or knowledge of Bayes nets, using such tools can be very helpful when organizing and thinking about the complicated interactions among examinee characteristics, test and administration features, and assessment purposes. It is expected that Bayes nets no more

complex than those shown in this report are likely to prove useful. At the same time, much is also likely to be gained from Bayes nets that are more complex than those shown.

Evidence-centered assessment design shows promise in supporting explicit structuring and representation of assessment arguments as they pertain to test takers with disabilities. While the models of these assessment arguments cannot mechanistically make key decisions about test design and use, they can illuminate the nature of decisions and help assessment designers think through the sometimes-competing goals of assessment designs. It is hoped that the approach will also help assessment planners think creatively about how to include individuals with a greater variety of disabilities by expanding the list of accommodations available for consideration that yield valid assessment results. Furthermore, a robust validity framework should also help determine the nature and scope of the features that are made available under the heading of universal designed assessments for a given audience and purpose.

***Recommendations for Further Work***

The preceding material represents a first step in showing that ECD can illuminate decisions associated with assessment accommodations for individuals with disabilities. Following are further steps that could be taken in preparing to apply this approach for the benefit of individuals in special populations taking NAEP and other large-scale assessments.

1. Work with NAEP assessment planners to attempt a more comprehensive mapping of the NAEP frameworks into ECD concepts and structures. Identify way of making the definitions of targeted proficiencies more explicit. Given the challenges encountered in doing this analysis, this seems among the most important steps.[114]

2. Expand the current analysis by addressing accommodations related to English-language learners (e.g., bilingual booklet, bilingual dictionary).

3. Focus more on the characteristics of populations by looking at the variety of characteristics of individuals constituting the population.[115] Give particular attention to the levels of specific KSAs that are most characteristic of particular disabilities, particularly invisible disabilities, such as learning disability. Survey research about the accuracy of diagnosis as well as the prevalence of different kinds of disabilities.

4. Identify ways of helping assessment planners use ECD to explore what-if analyses for decisions such as accommodation policies, possible universal design features, reconsideration of the definition of targeted proficiencies, and so on.[116]

5. Conduct research on important accommodations and access features such as extended time, calculator, read-aloud, and reduction in various kinds of linguistic load.

6. Develop models of utility, fairness, and feasibility to inform accommodation-related decisions. For example, one may determine that for an assessment of a given type, the harm caused by a false-positive outcome is less (or greater), overall, than the harm caused by a false-negative outcome. Since any assessment design is a matter of tradeoffs—accommodations that eliminate a source of invalidity for one kind of test taker may open the door to invalidity for another, or impose a high cost—designs about accommodation policies seem natural to address within utility-based analyses that allow examination of the consequences of test use.

7. Identify how processes such as test development processes and determination of accommodations can better support the knowledge acquisition needed to reason effectively about accommodations.

8. Explore ways of attaching backing (or support) to the warrants of the models. The warrants of the argument are implemented in the Bayes nets as conditional probabilities. Convenient ways need to be devised for attaching backing to the warrants so that substantive knowledge associated with an argument is not lost or ignored.

9. Explicate the strengths and limitations of the different criteria for judging the validity of accommodations. Consider especially those experimental designs that have individuals both with and without disabilities take tests with and without accommodations (Phillips, 1994; Fuchs, Fuchs, Eaton, Hamlett, & Karns, 2000; Tindal, Heath, Hollenbeck, Almond, & Harniss, 1998).

10. Explore ways that the validity argument framework can sharpen research hypotheses and otherwise guide empirical research, especially regarding invisible disabilities such as a learning disability.
11. Identify complementary ways to express rationales for accommodation-related decisions (e.g., narrative descriptions of the argument).

## References


Abedi, J. (2001). *Assessment accommodations for English language learners: Issues and recommendations* (Policy Brief No. 4). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J. (2002, December). *Accommodations for English language learners.* Paper presented at ETS, Princeton, NJ.

Abedi, J. (2003, April). *Language factors in the assessment of English language learners.* Training session presented at the annual meeting of the American Educational Research Association, Chicago.

Abedi, J., Courtney, M., & Leon, S. (2003). *Effectiveness and validity of accommodations for English language learners in large-scale assessments* (Center for the Study of Evaluation [CSE] Rep. No. 608.) Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., Leon, S., & Mirocha, J. (2001). *Impact of students' language background on standardized achievement test results: Analyses of extant data.* Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., Lord, C., & Plummer, J. R. (1997). *Language background as a variable in NAEP mathematics performance* (Tech. Rep. No. 429). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Aguirre-Muñoz, Z., & Baker, E. (1997). *Improving the equity and validity of assessment-based information systems* (CSE Tech. Rep. No. 462). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.

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 May 30, 2008, from http://www.access-board.gov/sec508/508standards.htm

Bachman, L. F. (1990). *Fundamental considerations in language testing*. New York: Oxford University Press.

Chisolm, W., Vanderheiden, G., & Jacobs, I. (Eds.). (1999). *Web content accessibility guidelines* (W3C recommendation). Retrieved May 30, 2008, 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). *The impact of providing testing accommodations to students with disabilities.* Paper presented at the annual meeting of the American Educational Research Association.

Embretson, S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93,* 179–197.

Frase, L. T., Chudorow, M., Almond, R. G., Burstein, J., Kukich, K., Mislevy, et al. (2003). Technology and assessment. In H. F. O'Neil & R. Perez (Eds.), *Technology applications in assessment: A learning view* (pp. 213–244). Mahwah, NJ: Erlbaum.

Fuchs, L. S., Fuchs, D., Eaton, S. B., Hamlett, C., & Karns, K. (2000). Supplementing teacher judgments about test accommodations with objective data sources. *School Psychology Review, 29*(1), 65–85.

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., & Mislevy, R. J. (in press). *Design patterns for improving accessibility for test takers with disabilities.* Princeton, NJ: ETS.

Hansen, E. G., Mislevy, R. J., & Steinberg, L. S. (2003, April). *Evidence-centered assessment Design and individuals with disabilities.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Hansen, E. G., Mislevy, R. J., Steinberg, L. S., & Forer, D. C. (2003, February). *Accommodating special needs candidates: Evidentiary foundations.* Paper presented at the annual meeting of the Association of Test Publishers (ATP), Amelia Island, Florida.

Hansen, E. G. & Steinberg, L. S. (2004). *Evidence centered assessment design for reasoning about testing accommodations in NAEP reading and math.* Unpublished paper commissioned by the Committee on Participation of English Language Learners and

Students with Disabilities in NAEP and Other Large-Scale Assessments of the Board on Testing and Assessment (BOTA) of the National Research Council.

Heath, A., & Hansen, E. G. (2002). Guidelines for testing and assessment. In IMS Global Learning Consortium (Ed.), *IMS guidelines for developing accessible learning applications* (section 9). Retrieved May 30, 2008, from http://www.imsproject.org/accessibility/accv1p0/imsacc_guidev1p0.html#1312344

IMS Global Learning Consortium. (2002). *IMS guidelines for developing accessible learning applications*. Retrieved May 30, 2008, from http://imsproject.org/accessibility/accv1p0/imsacc_guidev1p0.html

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: Erlbaum.

Lutkus, A. D., & Mazzeo, J. (2003). *Including special-needs students in the NAEP 1998 Reading assessment, Part I, comparison of overall results with and without accommodations* (NCES 2003-467). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics.

Markman, A. B. (1999). *Knowledge representation.* Mahwah, NJ: 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., Chudowsky, N., Draney, K., Fried, R., Gaffney, T., Haertel, G., et al. (2003). *Design patterns for assessing science inquiry* (Principled Assessment Designs for Inquiry [PADI] Tech. Rep. No. 1.) Menlo Park, CA: SRI International.

Mislevy, R. J., Sheehan, K. M., & Wingersky, M. S. (1993). Equating tests with little or no data. *Journal of Educational Measurement, 30,* 55–78.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1,* 3–67.

National Assessment Governing Board. (in press). *Mathematics framework for 2005: National Assessment of Educational Progress*. Washington DC: National Assessment Governing Board.

National Assessment Governing Board. (2002). *Reading framework for the 2003 National Assessment of Educational Progress.* Washington, DC: National Assessment Governing Board.

National Assessment Governing Board. (2003). Accommodations most frequently provided by NAEP. In *NAEP assessment coordinator manual* (pp. 4.42–4.44). Washington, DC: Author.

National Center on Educational Outcomes. (2007, March). *Special topic area: Universally designed assessments*. Retrieved May 30, 2008, 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.* 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.

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, Washington.

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

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.

Thompson, S., Thurlow, M., & Moore, M. (2002). *Using computer-based tests with students with disabilities* (Policy Directions No. 15). 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 Rep. No. 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Thorndike, E. L. (1971). Reading as reasoning: A study of mistakes in paragraph reading. *Reading Research Quarterly*, *6*(4), 323–332.

Tindal, G., Heath, B., Hollenbeck, K., Almond, P., & Harniss, M. (1998). Accommodating students with disabilities on large-scale tests: An experimental study. *Exceptional Children, 64*(4), 439–450.

Toulmin, S. E. (1958). *The uses of argument*. Cambridge, England: Cambridge University Press.

Wigmore, J. H. (1937). *The science of judicial proof* (3rd ed.). Boston: Little, Brown, & Co.

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., Ragosta, M., Bennett, R. E., Braun, H., Rock, D. A., & Powers, D. E. (Eds.). (1988). *Testing handicapped people*. Boston, MA: Allyn and Bacon.

## Notes

[1] "States regulations for excluding special-needs students from their assessments or for providing such students with a range of adaptations/accommodations do not necessarily match the guidelines issued by NAEP. . . One reason for this divergence is the slightly more limited range of adaptations/accommodations offered in NAEP compared to some state assessments. . . . Some accommodations, such as a helper reading aloud the reading passages in a reading test, are permitted in some state assessments, but not in the NAEP reading assessment" (U.S. Department of Education, 2003, p. 71).

[2] Assessment content could be read aloud by a live reader (e.g., administrator), generated and delivered via computer, or delivered in prerecorded audio mode. In NAEP, the read-aloud accommodation is delivered via live reader.

[3] NAEP provides several different methods for implementing the read-aloud accommodation, where required by Individualized Educational Plan (IEP) documentation for the student (National Assessment Governing Board, 2003):

1. Read-aloud in regular session—The student raises their hand if they need a word, phrase, or sentence read aloud. They use their regularly assigned assessment booklet.
2. Small group—This is for students requiring major portions of the booklet to be read aloud. All students in the group are assigned the same booklet.
3. One-on-one—In some cases, students may have the test read aloud to them in a one-on-one setting using their regularly assigned booklet.

[4] This is reiterated later in the same document, this time using the term *reading* instead of *reading comprehension*: "Because NAEP considers the domain of its reading assessment to be reading, the assessment cannot be read aloud" (2003 NAEP Reading Framework, p. 25).

[5] A tool for creating and editing Bayes nets may facilitate the process of adding evidence by providing a graphical user interface that allows the Bayes net user to select from a list or menu the desired level (or value) for any node (variable). Making the selection fixes the variable at whatever value has been designated and is considered to be adding evidence.

[6] As will be discussed, an output of the likely validity involves assumptions about the definition of validity in such contexts.

[7] We sometimes treat the term *targeted proficiency* as a synonym for *construct*, although we prefer the former term because of the narrower, more precise meaning that we have assigned.

[8] We provide a more technical definition of effective proficiency and its relationship to performance later in the report.

[9] The paper essentially describes and summarizes the June, 16 2003, presentation to the Committee on Participation of English-Language Learners and Students with Disabilities in NAEP and Other Large-Scale Assessments, augmenting it with some additional detail and analysis. Consistent with that presentation, this paper will give more attention to reading than to mathematics. Though the presentation did not address English language learners, the paper will discuss briefly a few issues related to accommodations for that population. On the other hand, science content, which received some attention in the presentation, is not addressed in this paper.

[10] The paper includes, by permission, a significant amount of copyrighted material of ETS, specifically, that which is being prepared for publication.

[11] Cognitive psychologists study two kinds of knowledge representations. External representations are artifacts and symbol systems that people construct, share, and work with in the world. Internal representations are *in peoples' heads*—patterns of information present in some form—the means and the subject of cognition. Our concern here is with external representations.

[12] For example, professional standards require that "unless evidence for validity for a given inference has been established for individuals with specific disabilities, test developers should issue cautionary statements in manuals or supplementary materials regarding confidence in interpretations based on such test scores" (Standard 10.4, American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME], 1999, p. 106).

[13] While not critical to this paper, it may be useful to some readers to have a somewhat more technical description of effective proficiency. We regard effective proficiency as a latent variable that is essentially like what is termed the *true score* in standard psychometrics (i.e., an expected score for performance on a test under the conditions it is administered and with

whatever nonfocal and focal requirements it imposes). As implied in its similarity to the true score of standard psychometrics, we are modeling effective proficiency as excluding measurement error. In this paper, we sometimes refer to effective proficiency as related to performance under operational conditions; this is a somewhat less precise convention, since we would normally think of performance as visible rather than hidden (latent). Despite this imprecision, this convention underscores the idea that a person's level of target proficiency is independent of operational conditions, whereas one's effective proficiency is closely related to actual performance in an operational setting.

[14] In a formal version of ECD, there are more relationships across layers than we discuss here.

[15] Inferential reasoning is sometimes called *common-sense reasoning*, but as Toulmin emphasized, the *process* of reasoning cannot be carried out in the absence of warrants, or rationales that justify the reasoning process. The everyday use of the term common-sense reasoning appears to encompass many assumptions about the way that the world—things, people, organizations, etc.—works.

[16] Notice that this warrant reasons deductively, from claim (good reading comprehension) to data (correct performance on the main idea item). This is consistent with the idea underlying psychometric theory that in building psychometric models, one reasons primarily in a deductive direction: If the student had such and such values of variables that characterize knowledge or skill, what would be the probability distributions for various behaviors they might produce, such as test scores and performances? On the other hand, during operational scoring of the assessment, one reasons in the inductive direction, from the data (an incorrect score) to the claim (that Sue has poor reading comprehension): Given a student's performance, what might be the values of the variables representing their knowledge or skill?

[17] "Most states have a list of possible or common accommodations for students with disabilities within the categories of presentation, response, timing/scheduling, and setting" (Thompson, Thurlow, & Moore, 2002).

[18] This situation brings to mind a principle mentioned in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999), which states that "an accommodation for a particular disability is inappropriate when the purpose of the test is to measure the presence and degree of that disability" (p. 102). Although the spelling test may not have been

explicitly intended to detect a spelling disability, the definition of the spelling construct as well as features of the content of the spelling test may be very similar to those of a test that is explicitly intended to diagnose spelling disability.

[19] A subtler example is coincidental familiarity with content topics of the test content; by chance the student may be much more familiar with the test content than his or her classmates. There may be little one can do to address this alternative explanation, other than to design tests that rely minimally on nonessential content knowledge and include tasks that span a variety of content situations. Coincidental reasons that a specific item will be harder for some people and easier for others, above and beyond their targeted proficiencies, are accounted for in measurement models such as item response theory (IRT). These are sources of variation between individual performances and the effective proficiencies tapped by a given testing configuration.

[20] However, the 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: (a) to facilitate participation by students with disabilities and (b) to increase the validity of scores" (McDonnell, McLaughlin, & Morison, 1997, p. 204). That approach recognizes the purpose of increasing participation as a potentially legitimate function of an accommodation. Recognizing multiple purposes for accommodations requires a richer way of representing the argument about what constitutes an appropriate accommodation. There are undoubtedly many instances in which inclusion can be increased without compromising validity, yet some potentially serious compromises can arise and one may need to decide what kinds and levels of inclusion would allow what kinds and degrees of compromise to particular facets of validity.

[21] In the examples used in this report, that level happens to be called good level.

[22] Our use of terms in intended to reflect a distinction between the definition of what is intended to be measured (which allows the distinction between focal and nonfocal KSAs) and the actual demands or requirements for examinee skills that occur in operational testing. Ideally, what is intended to be measured is what actually gets measured in the operational setting, but the issue of validity seems to hinge on the nature and magnitude of discrepancies between intent and actual implementation of an assessment.

[23] In this report we will focus on the appropriateness of the task features.

[24] What we refer to as nonfocal requirements (Haertel & Linn, 1996) seem to correspond closely to 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 argue, "If some examinees are deficient in a test's ancillary abilities, then it is biased against them" (1996, p. 63). They are similar to additional KSAs (Hansen & Mislevy, in press), which (in contrast to focal KSAs) are "other knowledge/skills/attributes that may be required in a task developed from a design pattern" (Mislevy et al., 2003). They also appear to have similarities to what are called *access skills* (Elliott & Roach, 2002, p. 12). However, a key point that we attempt to emphasize is that, as we use the terms, the distinction between nonfocal and focal KSAs is driven by the definition of the targeted proficiency while whether or not a KSA is requirement is driven by whether the KSA is necessary to perform well in an actual 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.

[25] For example, suppose that there is an nonfocal requirement that is very high (i.e., that a high level of a particular nonfocal KSA is necessary to do well in an operational setting). But if an assessor has ascertained prior to the test administration that everyone being tested can satisfy that nonfocal requirement, then it does not thwart inferences. However, in accommodation-related situations, especially, where difficulties in satisfying nonfocal requirements have not previously been diagnosed or do not qualify as disabilities, there is a need for a thorough examination.

[26] This recipe for valid assessment is simplified since, as noted elsewhere, focal requirements can be too high or too low, even where nonfocal requirements have been satisfied.

[27] The terms *requirement* and *demand* are sometimes used interchangeably, but have somewhat different connotations as used in this report. Both are ways of expressing the need for KSAs induced by features of the task (e.g., item) situation. We think that it may be worthwhile to suggest a subtle difference in connation. Specifically, the term demand connotes this need as it comes directly from the task feature and the term requirement connotes this need as it exists

further down the chain of reasoning. By the time the requirement is paired with an examinee ability, then the level may have been moderated by other influences.

[28] It should be noted the this report informally treats the terms *skill* and KSA essentially as synonymous, so that we use the terms focal skill and nonfocal skill as virtual synonyms for focal KSA and nonfocal KSA, respectively, even though the latter terms are more encompassing. The report also uses the terms *test* and *assessment* almost synonymously, even though the latter term is more encompassing.

[29] In some assessments of communication skills such as listening comprehension, a sensory skill—such as hearing—is a focal (as opposed to nonfocal) requirement. Yet even a listening comprehension test may have other sensory-related nonfocal requirements, such as requirements for sight to receive the items' visually displayed text and accompanying visuals, if applicable.

[30] Ideally, problems would be anticipated during test design and then addressed before or during the pre-administration phase of assessment system activity.

[31] There are other tools for representing assessment arguments, notably including Portal, an ECD-based tool developed by Mislevy, Steinberg, and Almond (Frase et al., 2003). The Portal program has rich facilities for representing the parts of the argument and for documenting the backing for warrants.

[32] These conditional probabilities are a part of the warrants of the Toulmin argument structure. In addition to the substantive grounding of the argument—the why—the conditional probabilities characterize the direction and the strength of the evidence—the which way and how much.

[33] Element two of the seven elements of universal design of assessments is *precisely defined constructs* (Thompson, Johnstone, & Thurlow, 2002).

[34] It generally seems appropriate to view these levels as exhaustive and mutually exclusive, though there are probably exceptions to that rule.

[35] Again, the point of this illustration is not to say how RC should be defined, but rather that it is possible to define it in several ways. One could add other KSAs, such as *recognize words*, *know English vocabulary*, etc. One could even add focal KSAs such as *know how to snow ski*.

The issue here is not what the most defensible definition of RC is or even what the definition of RC that is most amenable to certain accommodations is. The key issue is simply the idea of a structured way of defining constructs.

[36] Understanding this convention is very helpful in understand examples later in the report.

[37] The direction of arrows is a convention, but it does have some important semantics. Toulmin's arrow (and the ones in Wigmore's diagrams as well [Wigmore, 1937]) go from data to claim, since that is the direction that one reasons in individual cases. MSBNx and Portal Bayes nets fragments go from proficiency to the observable because this is the way the statistical/measurement model, effectively a warrant for reasoning in latent-variable models, is built. [38] How one comes up with a definition of the targeted proficiency is one of the most important issues in the proposed approach. This topic receives further discussion later in this report.

[39] Other possible variables might include meet demand for recording answers, meet demand for working quickly, etc. However, focusing on meet reception demand in this simple example seems appropriate, given that many testing accommodations are intended to overcome barriers to receiving test content (e.g., blind, low vision, deaf).

[40] The assumption that our model encompasses an exhaustive set of circumstances that could prevent RC from matching effective RC is a great convenience in understanding the behavior or the model.

[41] We will discuss later the probabilities associated with the item score node.

[42] Below each node (oval representing a variable) is a list of values that the variable can assume (in this case, good and poor for variable RC). The number (in parentheses) to the right of the value name is the probability (or strength of our belief) about the variable having that value. For example, there is a probability of 1.0 (i.e., 100%) that the RC has the value of good. The horizontal length of the color bar corresponds to the just-mentioned probability from 0.0 to 1.0. The color of the bar is irrelevant for the purpose of this discussion.

[43] One could argue that the probability of answering the item correctly should be zero rather than 20% as shown in this picture. This would be because the examinee cannot see the answer sheet to guess one of the five possible answers. However, this model is focusing strictly on

reception of test content and hence assumes that there is no barrier to the responding. In this case, even if reception demand is not met, the test taker would still be able to guess even when reception demand is not met. A more complete model might reflect that fact that sensory difficulties (e.g., blind) that create barriers to reception might also create a barrier to response (e.g., inability to use a mouse).

[44] Though poor RC and good, effective RC would clearly constitute a false-negative outcome, such an outcome is not possible given the conditional probabilities that we have coded into this particular Bayes net.

[45] In typical usage, the term construct change seems more commonly used in connection with apparent false-positives than with apparent false-negatives.

[46] Ideally, meeting the reception demand would require satisfying only nonfocal requirements (i.e., demands for nonfocal KSAs). Yet, as we will see later, focal requirements may also need to be satisfied in order to have reception demand met. Bayes net models can reflect such subtleties, if applicable.

[47] For the sake of clarity, the model has been made simple. Aspects of that simplicity include:

1. Use of 1 and 0 wherever possible, rather than probabilities such as 0.75 and 0.25
2. Few (usually two or three) states for variables
3. Focus on a single item rather than many
4. Mild (flat) prior probabilities on variables that have no parents. (For the variable RC, flat prior probabilities consist of 0.5 probabilities for good and 0.5 for poor.). These priors (prior probabilities) are the conditional probabilities that are entered by the model developer at the time of model creation. In the context of this report, priors for root nodes (nodes that have no parents) are essentially immaterial, since we generally use these nodes as decision variables that we set to a specific value and therefore the prior has no impact on the machinery of the argument.

[48] It should be noted that the symmetry between the 20/80 and 80/20 splits is not intended to imply a meaningful symmetry. For this discussion the values of the probabilities are almost immaterial so long as they are not 1 or 0. For a five-choice single-selection multiple-choice item, it does seems reasonable that given poor effective RC, the probability of a correct

answer should be about 20%, although since the poor category may encompass a range of low-end abilities, a higher%age would also be tenable. For good effective RC there would likely be a wider range of viable probabilities of a correct answer, for example, a range of 70 to 95% or wider. The key point is that an observation (score) does not provide us with perfect information about even effective proficiency.

[49] The conditional probabilities where see = yes suggest that a fully sighted person is equally able to receive content whether the font size is regular or large. This does not capture the nuance that reception of visually displayed text may depend, to some extent, on other factors, such as how much scrolling or page turning is required to access the item content.

[50] Some person variables could be influenced by test designers. For example, if there were a person variable called familiar with test format with possible values of yes and no, then a test designer could help ensure that adequate familiarization materials were made available beforehand to test takers.

[51] On the other hand, for assessments of sight and hearing, respectively, sight and hearing are the targets of measurement.

[52] The Portal tool provides facilities for storing and retrieving backing information in the form of short pieces of text as well as full documents.

[53] We could have collapsed the chain of reasoning to go from kind of item directly to meet RC demand (especially since we are modeling RC demand as dependent on only one variable—kind of item). Yet we have elaborated the chain of reasoning to emphasize the importance of the notion of "demand" — what skill the task or item requires of the test taker in order to succeed — in the chain of reasoning.

[54] Evidence-centered assessment design (ECD) has traditionally given considerable attention to the variables that drive demand for the targeted skill, such as the kind of item variable that drives demand for reading comprehension ability (RC). The current effort tends to additionally highlight demands for nonfocal KSAs (e.g., see), which are critical to an understanding of testing accommodations.

[55] The CAF is sometimes also called the Assessment Blueprint.

[56] This assumes, for example, that this is the proper setting for the same variable in the case of a nondisabled person whose targeted proficiency is validly measured under default conditions. This third criterion is arguably best viewed as dependent on the second criterion being met. The important issues of which validity criteria are most important and practical (e.g., ascertainable) for various testing purposes is largely beyond the scope of this paper.

[57] Neither should it increase the demand for focal skills. However, in testing accommodations, reduction in demand is generally considered the larger threat. The modeling approach is fully capable of dealing with either kind of threat. Bayes nets shown later in this paper are better suited to illustrating excessive focal demands (requirements).

[58] We are reasoning in a deductive direction, that is, based on specific values of the key parent variable (i.e., RC), and seeing what impact that has on the key child variable of interest (i.e.,effective RC), given a set of testing conditions.

[59] One may notice that it would be possible to have a true-negative outcome in which the poor performance (effective RC) could have resulted from one or both of two circumstances—poor RC and reception demand not being met. In Case 2, reception demand is met.

[60] We assume here that our reference group, consisting of nondisabled individuals who receive the test under default conditions, have their reading comprehension proficiency validity measured.

[61] As noted earlier, this model ignores the nuance that additional scrolling or page turning might make large font size a less effective alternative for an individual with unimpaired vision.

[62] This presence of ability pair 2 in this context is also consistent with the idea that the large-font accommodation would meet the differential boost criterion for valid accommodations (Phillips, 1994) (i.e., it boosts the performance of individuals with disabilities (low vision) but does not impact the performance of individuals without disabilities [nondisabled]). Satisfying this criterion also suggests that font size might reasonably be offered as a universal-design feature available to any test taker who desires it. A probably less-favorable approach would be to provide large font to all test takers. This particular model does not in itself guide the assessment planners between such alternatives, but models could be developed that incorporate knowledge about the appropriateness of such alternatives.

[63] The stated assumptions are most fully satisfied in the context of the examples diagrammed in this report when all items in the test are parallel to each other in terms of the focal and nonfocal demands that they place upon the test taker. This diagram is a simplification in the following sense. Suppose, as is often the case, that items are not parallel in terms of which and how much they require of various focal and nonfocal KSAs. One way of modeling the situation would be to map each item and its conditional probabilities given various focal and ancillary KSAs separately. The other, more common, approach, would be to posit a model at the level of test scores rather than test items, so the conditional probabilities would be composites across items with their own, not detailed, relationships with KSAs.

[64] It should be noted that a fully specified model addressing all possible nonfocal requirements can never be constructed, so one always confronts the issues of which and how many variables to explicitly include in an investigation.

[65] By accurate measurement, we mean that criteria even more stringent than those suggested earlier have been satisfied (e.g., fidelity to intent), all nonfocal requirements are appropriate and satisfied, and all focal requirements are appropriate (e.g., no reduction or increase from demand specified in definition of targeted proficiency).

[66] This specific discussion does not specifically involve individuals taking alternate assessments, since they were not considered eligible to take the assessment either under standard or accommodated conditions.

[67] There are some fairly prominent exceptions or violations of this assumption. For example, in subjects such as listening comprehension or speaking, individuals who are congenitally deaf would tend not to have the same ability distribution as a nondisabled subpopulation.

[68] Other issues that could be significant parts of a validity argument but which are not specifically addressed in this report include changes to reports, improved guidance in test score use, and the availability of alternate assessments.

[69] While the models discussed in this report do not explicitly reflect cost-oriented analyses of options, they do reflect subtler decisions that reflect an awareness about what is feasible and useful to explore, including relationships that are believed to have high utility for assessment

planners and about relationships between variables that are and are not cost-effective to explore.

[70] Other definitions of decoding could be provided by this one is sufficient for the purposes of this illustration.

[71] Braille is *not* on the list of frequently provided accommodations on NAEP assessments, but there is nothing to indicate that it may not have been provided on less frequent occasions (National Assessment Governing Board, 2003).

[72] For example, if decoding were not part of definition of the targeted proficiency, this would tend to encourage the use of read-aloud as an accommodation. Moreover, the change in definition could raise other issues. For example, would the test be fair to nondisabled test takers who receive no accommodation, yet must satisfy the decoding demand of the operational test? Perhaps the read-aloud feature could be offered to anyone who desired it, although such an approach might have significant cost implications. Furthermore, the test might become more difficult to differentiate from a test of listening comprehension. This is the kind of thorny issue that the approach raises yet also seeks to address in a principled way.

[73] A variety of factors that push for a richer representation of the argument are described in greater detail in Appendix B.

[74] In the case of Sue and low vision, there was only one focal KSA in the targeted proficiency so in that case the distinction between focal KSA and targeted proficiency is moot.

[75] These unintended effects thus account for some of the additional arrows (arcs) seen in typical Bayes net graphics.

[76] We used this shortcut in the case of Sue and low vision in going directly from font size to meet reception demand without going to a node called, let us say, demand for sight.

[77] Studies that evaluate the validity of accommodations sometimes involve experimental designs that have individuals both with and without disabilities take tests with and without accommodations (Almond, & Harniss, 1998; Fuchs, Fuchs, Eaton, Hamlett, & Karns, 2000; Phillips, 1994; Tindal, Heath, Hollenbeck,). While the specific criteria for accommodation validity vary among studies, the ideal situation for a valid accommodation is generally that individuals with disabilities gain a considerable performance benefit from the accommodation

but that individuals without disabilities gain no benefit from the same accommodation. In principle, the use of large font size is an accommodation would meet this criterion since large font size benefits individuals with low vision (relative to their performance with regular font size) but not individuals with unimpaired vision, since the latter will perform no better using large font size than they will with regular font size. Our Bayes net models are generally consistent with the logic of this criterion for evaluating validity of accommodations as well as with its application in the case of Sue and low vision, yet because we model RC and effective RC as having only two performance levels (good, poor) instead of three (good, okay, poor) we a have ceiling effect that prevents us from demonstrating this consistency as convincingly as we would prefer. Bayes nets could be built that have the requisite number of levels for demonstrating the operation of this criterion. As one might surmise, the differential boost criterion for accommodation validity has limitations and some researchers have advised against using it as sole criterion for determining the appropriateness of an accommodation (Elliott & Roach, 2002, p. 11–12). The domain modeling work discussed in this report seems capable of providing a framework for interpretations arising from such studies of accommodations. For example, such a framework could help make explicit conditions under which an observation of differential boost is most likely to be an indicator of an appropriate accommodation.

[78] Some KSAs seem physical or sensory in nature (e.g., see, hear, walk, while others are more cognitive and tend to be less visible [e.g., comprehend, decode]).

[79] A key example of such an intervening variable is the meet reception demand variable that indicates (yes or no) whether the person is able to receive the content. The direct parents of meet reception demand are other intervening variables (receive via visual text, receive via braille, and receive via audio, and presentation mode). These intervening variables (and the conditional probabilities) reflect a psycho-physical model that says, in brief, that if content is presented in a particular presentation mode (e.g., visual text, braille, or read-aloud) and if the value of the corresponding modality-specific reception variable (e.g., receive via visual text, receive via braille, and receive via audio) is yes, then the person can receive the test content.

[80] In a sense, the distinction that the assessment planner makes between focal KSAs and nonfocal KSAs when creating the psycho-physical model is likely to be helpful in domain modeling.

Ideally, the psycho-physical processes related to focal KSAs would be separable from those related to nonfocal KSAs. For example, ideally, the meet reception demand node would be affected only by nonfocal KSAs, such as when we modeled meet reception demand depending on the proper match-up of a single nonfocal KSA and a sign task model variable (font size). Examples where this ideal was not fulfilled include cases in which meet reception demand would be affected by the focal KSA of decoding (using Definition A) in addition to the nonfocal KSA of sight, because the presentation mode is visual text.

[81] It is possible that some of the validity issues encountered in our later discussion of the regular dictionary would also arise in connection with the bilingual dictionary.

[82] On the other hand, the NAEP framework provides additional detail in other areas that are not the focus of this investigation.

[83] Evidence for NAEP reading as involving *English* was found in a separate document (U.S. Department of Education, 2003). This would have been important information in an ECD design for addressing the needs of English-language learners.

[84] The relationships would undoubtedly be probabilistic, and a measurement model could be introduced in which the parameters of the NAEP IRT models were conditioned on these complexity variables (see, for example, Mislevy, Sheehan, & Wingersky, 1993).

[85] Option 2 seems to be elaborated on by the following language: "The demands on thinking that an item makes—what it asks the student to recall, understand, reason about, and do—are determined based on the assumption that the student is familiar with the *mathematics of the task*. If a student has not studied these mathematics, the task is likely to make different and heavier demands, and the student may well not be successful on it" (National Assessment Governing Board, 2001, page 18, emphasis added). It is not clear how to interpret the reference to the "assumption that the student is familiar with the *mathematics of the task*" (emphasis added). This statement begs more specific questions about how are the following terms related to each other: familiarity with the mathematics of the test, math complexity, math ability, knowledge of math vocabulary.

[86] It is worth noting that examination of the validity of accommodations may cause us to refine and extend our definition of the targeted proficiency.

[87] The key point is that both comprehension and reasoning serve as a dominant focal KSA in their respective assessments. However, it may be worth noting that there are also some arguments for a deeper similarity between comprehension and reasoning. For example, Thorndike (1971) argued for a view of reading as reasoning. Furthermore, reasoning is mentioned as being applicable to all NAEP questions, presumably regardless of subject area: "All NAEP questions emphasize critical thinking skills and *reasoning* rather than factual recall" (2003 NAEP reading framework, p. 20, emphasis added). Arguably, comprehension might be considered an aspect or subcategory of reasoning.

[88] This is a relatively simple set of cognitive skills that might be part of the definition of reading and mathematics. A more comprehensive list of constituent skills for reading might include, for example, the skills such as recognize words. Or the list of skills for mathematics might include KSAs such as know conventions for mathematical expressions (e.g., equations), perform basic math computations mentally, operate a four-function calculator, and know conventions for the display of charts or graphs. We trust, however, that even this short list of cognitive skills will serve to illustrate the basic approach.

[89] Recall that when a KSA is listed as n/a it means essentially the same thing as the lowest level of a KSA being the minimum level. By convention, any such KSA is defined as a nonfocal KSA and is therefore irrelevant to the definition of the targeted proficiency.

[90] The model also has a third definition (alternative mathematics) that will be discussed later.

[91] The role of the dictionary will be discussed later in detail.

[92] Recall that our convention was that any KSA for which the lowest level of the KSA was sufficient was defined as an nonfocal KSA.

[93] Essentially the same statement can be made for focal KSAs: A focal KSA may or may not be required for good effective proficiency, since a poorly designed assessment may not require a focal KSA in order to perform well. We won't be discussing focal requirements as much as nonfocal requirements, since this focal requirements are an issue that is relevant for all examinees. The issue of focal requirements includes the importance of avoiding focal requirements that are too high or too low.

[94] Note that the Bayes net is capable of representing a wide range of situations, some in which the results of assessment would be invalid and others in which the results would be valid.

[95] One way to think about the value of being able to model 100,000-or-more situations is that as evidence is added to the model, such a model is capable of explicitly ruling out a very large number of alternative explanations.

[96] We have operationally defined dyslexia as being characterized by poor decoding ability. This is a very simple definition of dyslexia.

[97] Such an approach would involve a somewhat different use of diagnostic information than seems to usually be the case. Typically, diagnostic assessment of skills would be performed ahead of time to show deficits that point to the need for accommodations. However, such information could also be used to show strong capacities (absence of deficit) and that thereby a person would not be unfairly advantaged by an accommodation.

[98] For example, we somewhat arbitrarily built the model based on the assumption that decoding was part of the targeted proficiency (i.e., was a focal KSA) for NAEP reading, even though the framework did not actually state that it was.

[99] In the conclusion of this report, we also mention the possible value of the approach in informing and guiding empirical research.

[100] It is important not to overrely on a single criterion measure and to avoid overinterpreting the significance of either good or poor prediction of any one criterion measure. An example of the caution needed is exemplified by Willingham et al. (1988, p. 173): "The fact that a group is less predictable has never been a basis for identifying individual scores as being of doubtful comparability. It is even possible that if scores were disaggregated on the basis of other factors (especially the nature of the college program), they might prove *more* [emphasis added] valid than is typically true for nonhandicapped [*sic*] students! For example, if grading standards vary in two departments, it is readily demonstrable that pooled data are likely to yield lower validity coefficients than those found in each department separately."

[101] There may also be a problem with the criterion (e.g., contamination of the criterion).

[102] Another example of how research could inform model refinement concerns the adequacy and equivalence of different read-aloud methods. In this is report we have assumed that all read-

aloud methods (e.g., live reader, synthesized speech, and prerecorded audio) are essentially equivalent. However, nonequivalencies could arise. For example, consider the process that might be used for creating a prerecorded audio version of the read-aloud accommodation for a mathematics assessment. This would involve the creation of a script for the human reader to read aloud. The script would include a carefully crafted text description of each math *graphic* that is clear and understandable but does not give away the right answer. Once recorded, this version of read-aloud could be delivered via CD-ROM, audiocassette, or other means. Now consider, by contrast, a human reader who is not provided with a text description of graphics, has little or no time to study the exam before reading it aloud, and must quickly judge on-the-fly how to effectively describe the math graphics without giving away the answer (Ruth Loew, personal communication, April 2, 2004). Empirical examination of these two different methods for delivering read-aloud could inform refinement of the model. (Obviously, it could also inform the development and implementation of procedures for ensuring that a live reader has access to a carefully crafted description of mathematics graphics.)

[103] The typical working assumption is that the linguistic demand of a reading assessment is intentional. While it is worthwhile to examine this assumption, it is well enough accepted that most efforts to reduce linguistic demand naturally focus on nonreading domains.

[104] As noted earlier, examples of possible content vocabulary found in the sample items it in the 2005 NAEP mathematics framework would be words like cube (p. 61), vertex (p. 61), etc.

[105] The accuracy of existing disability classifications is a significant issue in accommodation decisions.

[106] The World Wide Web Consortium (W3C) *Web Content Accessibility Guidelines* (Chisolm, Vanderheiden, & Jacobs, 1999), the *Electronic and Information Technology Accessibility Standards (Section 508*) (Architectural and Transportation Barriers Compliance Board, 2000), and the IMS *Guidelines for Developing Accessible Learning Applications* (IMS Global Learning Consortium, 2002) are useful points of reference for accessibility guidelines and standards (see also Heath & Hansen [2002]).

[107] List item “adequate empirical relationship between accommodated scores and performance in criterion situations ” could alternatively be framed as part of Step 5 on verification of results.

[108] One heuristic in this process is to think about or examine diverse criterion situations (including those involving individuals with disabilities using normal life-style accommodations, if appropriate) and determining the level of the KSA needed to perform well, given good proficiency in each of the other focal KSAs.

[109] One can envision other kinds of combinations, but the conjunction seems the simplest to think about and implement.

[110] Innovative accommodations can require an assessment planner to think imaginatively about how to parse the full set of person characteristics into KSAs; such an exercise can split, merge, or otherwise modify definitions of nonfocal KSAs, thereby reducing or eliminating some nonfocal requirements.

[111] Recent interest in the concept of universally-designed assessments underscores the need for such a validity framework. Universal design is intended to make a wide range of accessibility features available to individuals who need or desire them. According to Thompson and Thurlow (2002): "The need that many students have for accommodations could be reduced if assessments could be universally designed." An implication of the term *universal design* is that while an accommodation is an accessibility feature that generally requires prior approval, a universal-design feature is an accessibility feature that may be made available as needed or desired by the test taker, without prior approval. For example, if the feature that increases the size of fonts on a computer screen is considered a universal design feature then it is made available to any individual who perceives a need for it. As an aside, doing this would increase the need for scrolling, which might induce other nonfocal requirements. A robust validity framework would help determine the nature and scope of the universality that is feasible within an assessment design for a given audience and purpose.

[112] Consideration of a change in the definition of the construct can, if nothing else, help clarify the original definition.

[113] Indeed, our main emphasis relative to focal KSAs has been on the importance of avoiding accommodations that change, especially decrease, focal requirements from their intended levels. But we also want to avoid excessive focal requirements. For example, consider a mathematics targeted proficiency that defines good mathematics proficiency as consisting of two focal KSAs—good reasoning ability and okay knowledge of mathematics vocabulary.

The poorly designed test for that targeted proficiency might use excessively difficult mathematics vocabulary that requires a good (instead of okay) level of knowledge of mathematics vocabulary. Such an excessive focal requirement for knowledge of mathematics vocabulary would tend to put an individual with okay knowledge of mathematics vocabulary at a disadvantage.

[114] It should be noted that while difficulty was encountered in mapping from the framework documents to this application of ECD, other specific applications may not encounter such difficulties.

[115] This would be facilitated by developing ways of running scenarios automatically in batches rather than one at a time manually.

[116] An example of a fine-tuning of the definition of targeted proficiency would include greater specificity regarding what it means to do well on an assessment or to possess a good level of the targeted proficiency. This determination relates to how one distinguishes between focal and nonfocal KSAs

[117] In some implementations of the traditional ECD domain modeling, the interaction between nodes of the model is somewhat loosely defined—involving different kinds of associations between nodes, but not using specific logical/statistical relationships between nodes. The specification of rather precise logical/statistical relationships may be considered a key feature of the approach to domain modeling demonstrated in this report.

[118] Bachman (1990) notes: “All language tests must be based on a clear definition of language abilities, whether this derives from a language teaching syllabus or a general theory of language ability. As simplistic as this statement may seem, it turns out that designing a language test is a rather complex undertaking, in which we are attempting to measure abilities that are not very precisely defined, and using methods of elicitation that themselves, depend upon the very abilities we want to measure. This is the fundamental dilemma of language testing mentioned above: the tools we use to observe language ability are themselves manifestations of language ability” (p. 9). In other words, where the methods of elicitation rely on the same skills that we want to measure can make it very difficult to cleanly separate nonfocal from focal requirements.

[119] For example, suppose that in a reading comprehension test, decoding is not part of the intent of measurement but that the actual decoding load is excessive (i.e., higher-than-intended) for nondisabled test takers receiving the test in default conditions, which entails presenting content via visual text. Providing a read-aloud accommodation for a test taker who is blind seems reasonable, since doing so will allow reception demand to be met as well as tend to yield true (true-positive or true negative) outcomes (based on a comparison the targeted proficiency and effective proficiency). Yet it raises a fairness issues relative to many of the nondisabled takers who, let us assume, struggle with the excessive decoding load when taking the test under nonaccommodated conditions, and thus violating our assumption that individuals without disabilities are validly assessed under default conditions. Why should the disabled taker receive an accommodation but not the nondisabled students who struggle with the excessive decoding load?

[120] Such a situation can give rise to false-negative outcomes. In principle and by definition, the target of measurement does not include nonfocal KSAs. If there are any individuals in the target population whose nonfocal abilities do not satisfy the requirements for those skills, then this negatively impacts the validity of the assessment.

[121] Having decoding demand driven by this single variable is a considerable oversimplification. Many variables are likely to influence decoding demand. The use of this single variable is illustrative.

[122] Even an item that would have a high decoding demand when administered via visual text or braille is modeled as causing essentially no decoding demand when administered via read-aloud.

## Appendix A
## Evidence-Centered Design and the Current Effort

The approach in this report takes into account a range of variables that is wider than those in the ECD conceptual assessment framework (CAF). For example, this approach encompasses a larger range of person variables, notably, KSAs that are nonfocal in the sense of not being part of the intent of measurement. These nonfocal KSAs are not part of the CAF's student model, yet they can be a legitimate part of an ECD domain model. Nonfocal, but necessary, KSAs—nonfocal requirements—have not often been emphasized in traditional ECD-based designs.[117]

Figure A1 shows layers of assessment design and delivery. Layer D concerns the operation of an assessment. Some activities such as registration and familiarization occur in the pre-administration phase of assessment system activity; these are depicted as D1. The administration phase of assessment system activity, D2, may be what is most commonly associated with the test. This is where scores, items, and assessment conditions are quite visible, and where most of the tangible elements of accommodation policies appear. Events during the post-administration phase of assessment system activity, D3, entail activities such as score interpretation, which may call for additional information or additional inference in order to achieve the test's purpose. It is the higher layers in the system, however, that impart meaning to data that are obtained from the test. Layer A, for example, begins with the considerations of what the test is supposed to do, with what kinds of students, in what domain, under what constraints, and with what resources. This is called domain analysis in the ECD framework. Layer B, which is called domain model, concerns organizing this information in terms of an assessment argument. Toulmin diagrams, denoted in B2, are one representational form to assist the assessment designer at this stage. The representations—called paradigms in the ECD framework, labeled B1—are also useful. The idea of this layer is to sort, assemble, and organize the information obtained in domain analysis in a direction that will lead to design elements for an operational assessment.

Layer C, the CAF, is a formalization of design elements for the operational assessment, particularly those of the administration phase of assessment system activity. It is here that technical elements such as the variables and statistical forms of psychometric models are mapped out, rubrics or scoring rules are determined, work product requirements are specified, and essential task features are identified and codified to support task construction.

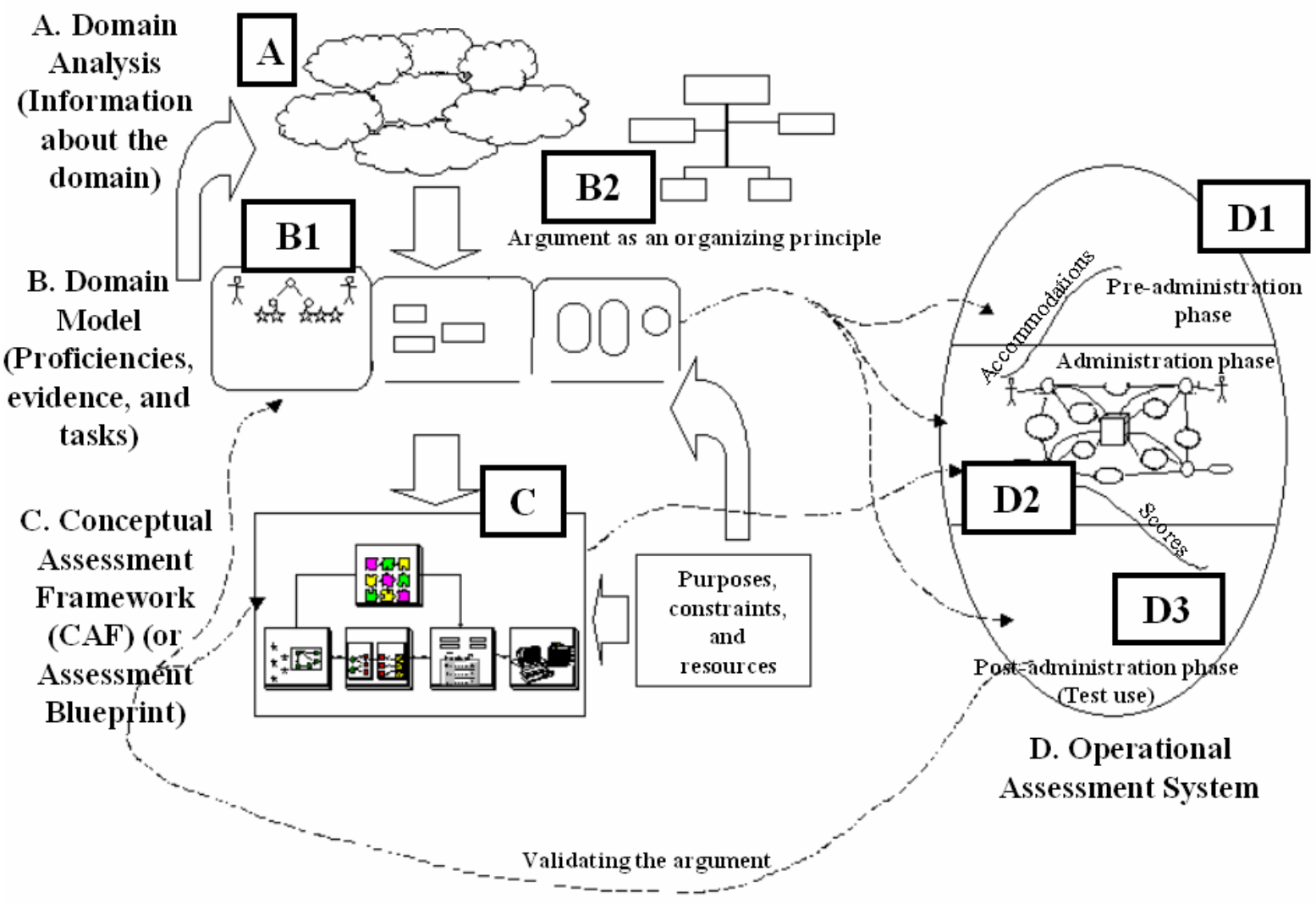


***Figure A1.*** **Layers in assessment design and delivery.**

## Appendix B

## Situations that Call for a Richer Representation

Several factors can make situations more complex, thereby calling for richer Bayes net representations of the assessment argument, beyond the simple Bayes net that we used for Sue and low vision.

- Greater diversity of nonfocal KSAs among the target population. Instead of a single nonfocal KSA (i.e., see) there may need to be many nonfocal KSAs in the model. The greater the importance of allowing all students (regardless of disability) to participate in an assessment, the greater the number of nonfocal KSAs that need to be considered.
- More task model variables. Instead of two task model variables (i.e., font size, kind of item) there may need to be many. A change in one task variable can dramatically affect demands for skills that are either nonfocal or focal.
- Complex target of measurement. Rather than being able to represent the targeted proficiency as a single node, multiple related nodes may be required. Typically, the greater the diversity of the test taker population, the greater complexity and precision that is needed in the definition of the focal and nonfocal KSAs.
- Processes that merge both focal and nonfocal requirements. For example, where decoding is part of the targeted proficiency (i.e., a focal KSA), nodes that one would like to indicate whether nonfocal requirements have been satisfied may also be influenced by focal requirements. For example, a node that one would like to characterize as pertaining only to nonfocal requirements (e.g., meet reception demand), may also include focal requirements (e.g. decoding [when that is part of the targeted proficiency]), since receiving content via braille or visual text depends on decoding skill. Such complexities often arise in assessment of language or reading-related skills.[118]
- Assessment products that are new, changing, or for which there is little body of knowledge about its underlying constructs or its uses.

- Complicated fairness issues. For example, reasoning about testing accommodations is generally easiest if one knows that test takers who are not members of special populations (e.g., nondisabled, native speaker of English) are validly assessed under default testing conditions. If that condition does not hold, then additional complications arise.[119]

## Appendix C

## Features and Benefits of the Richer Representation

This section describes the richer Bayes net representation of Tim and read-aloud accommodation. This Bayes net extends the capability of the net used for Sue and low vision in several ways.

1. Provides a richer representation of the intent of measurement. Instead of representing reading comprehension as simple single node (RC), it adds two nodes representing its component skills—comprehend and decode. Comprehend has two levels (good, poor) and decode has three levels (good, okay, poor).

2. Allows the user of the Bayes net to switch between different definitions of the targeted proficiency (see Table C1). As can be seen in Table C1 (conditional probabilities for the RC node), for example, under Definition A, in order for the person to have good reading comprehension (RC = good) they must have both good comprehension (comprehend = good) and okay or better decoding (either decode = okay or decode = good). On the other hand, under Definition B, in order for the person to have good reading comprehension (RC = good) they must only have good comprehension (comprehend = good) and their level of decoding ability is irrelevant. Thus, the key difference between the two definitions is manifest in what occurs with a person with good comprehension (comprehend = good) and poor decoding (decode = poor); while that individual is considered as having poor reading comprehension (RC = poor) under Definition A, they are considered as having good reading comprehension (RC = good) under Definition B.

3. Expands the ranges of the KSAs—especially nonfocal KSAs—to which we pay attention. Thus, instead of paying attention only to the skill of sight (see), this extended Bayes net includes the KSAs of decoding (decode), hearing (hear), touch (feel), as well as knowledge of braille codes (know braille codes). This wider range of KSAs greatly expands the richness of the model, allowing it to address, for example, situations involving additional (and multiple) disabilities and additional ways of receiving content (e.g., read-aloud, braille).

**Table C1**

***Conditional Probabilities for Reading Comprehension Based on Comprehension, Decoding, and the Definition of RC***

| Parent node(s) | | | RC | |
|---|---|---|---|---|
| Comprehend | Decode | Definition of RC | Good | Poor |
| Good | Good | A | 1.0 | 0.0 |
| | | B | 1.0 | 0.0 |
| | Okay | A | 1.0 | 0.0 |
| | | B | 1.0 | 0.0 |
| | Poor | A | 0.0 | 1.0 |
| | | B | 1.0 | 0.0 |
| Poor | Good | A | 0.0 | 1.0 |
| | | B | 0.0 | 1.0 |
| | Okay | A | 0.0 | 1.0 |
| | | B | 0.0 | 1.0 |
| | Poor | A | 0.0 | 1.0 |
| | | B | 0.0 | 1.0 |

*Note.* RC = reading comprehension.

This same information is formatted a bit more simply in Table C2.

4. Recognizes decoding as playing a potential role as either a nonfocal KSA or focal KSA, depending on the definition of reading comprehension that is used. For example, decoding can be a nonfocal KSA under Definition B or a focal KSA under Definition A.

5. Recognizes the importance of controlling the actual demand as well as the intended demand of tasks. The Bayes net attempts to take into account situations in which the

intended and actual demand for targeted or nonfocal skills diverge. For example, under Definition B, the intended demand for the decoding KSA may be null or low yet the actual demand for that KSA may be much higher.[120]

**Table C2**

***The Status of Reading Comprehension Based on the Learner State and the Definition***

| Learner state | Constituent skills | | Status of RC under definition: | |
|---|---|---|---|---|
| | Comprehend | Decode | A | B |
| 1 | Good | Good | Good | Good |
| 2 | Good | Okay | Good | Good |
| 3 | Good | Poor | **Poor** | **Good** |
| 4 | Poor | Good | Poor | Poor |
| 5 | Poor | Okay | Poor | Poor |
| 6 | Poor | Poor | Poor | Poor |

*Note.* Areas of difference in learner State 3 are in bold. RC = reading comprehension.

6. Recognizes how accommodations can impact demand for focal KSAs. In the case of Sue, the use of large font size did not impact demand for targeted skills. On the other hand, sometimes an accommodation can impact demand for targeted skills, as when a read-aloud accommodation reduces impact demand for decoding, which is a focal KSA under Definition A.

7. Provides a wider array of task model variables. In the case of Sue, there were two task model variables—font size and kind of item. In this Bayes net, we have added two new variables. First, the presentation mode variable has three values (visually text, read-aloud, and braille) and indicates how the test content is presented to the test taker. Second is the frequency of words with multiple contiguous consonants (FrqWdMultContigCons), which has three values (high, medium, and low). A word with multiple contiguous consonants such as strength is typically more difficult to decode than is a word of the same length but with fewer or no contiguous consonants (e.g., bananas). The high, medium, and low values of this variable, along with the

presentation mode (visual text, braille, read-aloud) drive, respectively, the high, medium, and low levels of decoding demand.[121, 122]

8. Recognizes how meeting decoding demand (meet decoding demand) for the purpose of receiving content depends on the person's decoding ability (decode) and the decoding demand. For example, according to the model, if a person has okay decoding ability, then they will be able to meet decoding demand if decoding demand is medium or low.

In summary, the extended application of the approach greatly increases the variety of situations that can be addressed. For example, with regard to person characteristics, instead of only addressing the requirements of individuals with low vision or who are nondisabled, the extended Bayes net can illuminate situations involving individuals who are blind, deaf, deaf-blind, or dyslexic.

## Appendix D

## Detail on Findings

**Table D1**

***Detail for Table 23***

| Name | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | R | M | R | R | M | R | M | R | M | R | R | M |
| Variable type = person | | | | | | | | | | | | |
| Comprehend, reason[a] | Good | Good | Good | Good | Good | Good | Good | Good | Good | Good | Good | Good |
| See | Yes | Yes | Partial | Partial | Partial | No | No | Yes | Yes | No | No | No |
| Hear | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Decode | Okay | Okay | Okay | Poor | Poor | Poor | Poor | Poor | Poor | Okay | Poor | Poor |
| Know braille codes | No | No | No | No | No | No | No | No | No | Yes | Yes | Yes |
| Feel | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Know content vocabulary | Poor | Okay | Poor | Poor | Okay | Poor | Okay | Poor | Okay | Good | Good | Okay |

*(Table continues)*

Table D1 (continued)

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Name | R | M | R | R | M | R | M | R | M | R | R | M |
| Know noncontent vocabulary | Poor | Poor | Poor | Poor | Poor | Poor | Poor | Poor | Poor | Poor | Poor | Poor |
| | Variable type = task model | | | | | | | | | | | |
| Presentation mode | Visual text | Visual text | Visual text | Visual text | Visual text | Read-aloud | Read-aloud | Read-aloud | Read-aloud | Braille | Braille | Braille |
| Comprehension/ reasoning driver | High | High | High | High | High | High | High | High | High | High | High | High |
| Decoding driver | Medium | Low | Medium | Medium | Low | Medium | Low | Medium | Low | Medium | Medium | Low |
| Content vocabulary driver | Low | Medium | Low | Low | Medium | Low | Medium | Low | Low | Low | Low | Medium |
| Noncontent vocabulary driver | Low | Low | Low | Low | Low | Low | Low | Low | Low | Low | Low | Low |

*(Table continues)*

Table D1 (continued)

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Name | R | M | R | R | M | R | M | R | M | R | R | M |
| Dictionary | None | None | None | None | None | None | None | None | None | None | None | None |
| Font size | Regular | Regular | Large | Large | Large | (Any) | (Any) | Regular | Regular | (Any) | (Any) | (Any) |
| | Decision = derived | | | | | | | | | | | |
| Definition of targeted proficiency | R | M | R | R | M | R | M | R | M | R | R | M |
| Targeted proficiency | Good | Good | Good | Poor | Good | Poor | Good | Poor | Good | Good | Poor | Good |
| Effective proficiency | Good | Good | Good | Poor | Good | Good | Good | Good | Good | Good | Good | Good |
| Valid[b] | Valid | Valid | Valid | Invalid | Valid | Invalid | Valid | Invalid | Valid | Valid | Invalid | Valid |
| Outcome = true | True-positive | True - positive | True - positive | True - negative | True positive | False-positive | True-positive | False-positive | True-positive | True-positive | True - negative | True-positive |
| Meet reception demand = yes | Yes | Yes | Yes | No | Yes | Yes | Yes | Yes | Yes | Yes | No | Yes |

*(Table continues)*

Table D1 (continued)

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Name | R | M | R | R | M | R | M | R | M | R | R | M |
| Focal requirements appropriate [c] | Yes | Yes | Yes | n/a | Yes | No | Yes | No | Yes | Yes | n/a | Yes |

*Note.* M = mathematics, R = reading, RC = reading comprehension.

[a] Comprehend refers to reading comprehension and reason refers to mathematics. [b] Valid = true outcome + meet reception demand = yes + focal requirements appropriate. [c] Neither too high nor too low; assumes reception demand is met.