*Educational Research and Evaluation*, 2013
Vol. 19, Nos. 2–3, 121–140, http://dx.doi.org/10.1080/13803611.2013.767614

Routledge
Taylor & Francis Group

# A "conditional" sense of fairness in assessment

Robert J. Mislevy[a]*, Geneva Haertel[b], Britte H. Cheng[b], Liliana Ructtinger[b], Angela DeBarger[b], Elizabeth Murray[c], David Rose[c], Jenna Gravel[d], Alexis M. Colker[e], Daisy Rutstein[b] and Terry Vendlinski[b]

[a]*Educational Testing Service, Princeton, NJ, USA;* [b]*Center for Technology in Learning, SRI International, Menlo Park, CA, USA;* [c]*CAST, Wakefield, MA, USA;* [d]*Graduate School of Education, Harvard University, Cambridge, MA, USA;* [e]*Independent Consultant, San Mateo, CA, USA*

Standardizing aspects of assessments has long been recognized as a tactic to help make evaluations of examinees fair. It reduces variation in irrelevant aspects of testing procedures that could advantage some examinees and disadvantage others. However, recent attention to making assessment accessible to a more diverse population of students highlights situations in which making tests identical for all examinees can make a testing procedure less fair: Equivalent surface conditions may not provide equivalent evidence about examinees. Although testing accommodations are by now standard practice in most large-scale testing programmes, for the most part these practices lie outside formal educational measurement theory. This article builds on recent research in universal design for learning (UDL), assessment design, and psychometrics to lay out the rationale for inference that is conditional on matching examinees with principled variations of an assessment so as to reduce construct-irrelevant demands. The present focus is assessment for special populations, but it is argued that the principles apply more broadly.

**Keywords:** evidence-centered design; diagnostic model; universal design for learning

## Introduction

Standardizing aspects of an examination process across examinees can reduce variations that would advantage some examinees and disadvantage others. Scores on exams under which, unbeknown to the score user, some examinees had more time than others or had their work rated by different criteria, for example, are patently unfair. *Unidentified nonequivalent surface conditions provide nonequivalent evidence about learners.* Ensuring that test materials and procedures are the same for all examinees epitomizes one sense of "fairness": All examinees run the same race, so to speak. Some particular remaining aspects, such as the content of specific test items, may favour some students and other aspects may favour other students, but the idea is that these are random differences which tend to average out (Green, 1978). We refer to this strategy as *marginal inference.* Marginal is a statistical term that means "averaging over".

Efforts to extend educational experiences to a more diverse population of students reveal that the same situation need not provide the same learning opportunities to all students. Similarly, the same assessment tasks may not produce the same information about what they know and can do. If we want to assess students' proficiency with arithmetic word problems, the same printed test may serve the purpose for a sighted student but not

*Corresponding author. Email: rmislevy@ets.org

© Educational Testing Service Inc. 2013. All Rights Reserved.

one with limited vision. *Equivalent surface conditions may not provide equivalent evidence about learners.*

Alternative forms of assessment such as accommodated tests, customized tests, and examinee-choice of tasks suggest a different sense of fairness: Tests can differ in their surface characteristics in such ways that equivalent evidence about examinees' proficiencies can be obtained (Rose, Murray, & Gravel, 2012). We refer to this as *conditional inference.* Conditional is also a statistical term, which means taking certain information into account specifically rather than averaging over the ways it might vary. In assessment, conditional inference means deliberately varying aspects of an assessment for students to enable each student to access, interact with, and provide responses to tasks in ways that present minimal difficulty, so the primary challenge is the proficiency meant to be assessed. Thus, *surface conditions that differ in principled ways for different learners can provide equivalent evidence.*

Assessment that is tailored in some form has become widespread, such as the accommodations spurred by the Americans with Disabilities Act. However, the methodologies of educational assessment and educational measurement (psychometrics) evolved in the environment of standardized assessment procedures and marginal statistical inference. Currently, much applied work with testing accommodations is after-the-fact: Unitary forms of tasks from standardized tests are first created, then retro-fitted in an ad hoc manner. We describe a prospective framework for coordinated design and analysis that supports conditional inference across tailored forms of tasks. We build on developments in three distinct areas that are required jointly, namely, assessment design theory, universal design for learning (UDL; Rose & Meyer, 2002; Rose, Meyer, & Hitchcock, 2005), and psychometric modelling.

The following section reviews the assessment-argument structure in which assessment design takes place, using an "evidence centered design" framework (ECD; Mislevy & Haertel, 2006; Mislevy, Steinberg, & Almond, 2003). The interplay among construct-relevant and construct-irrelevant knowledge, skills, and abilities (KSAs) with task features and work products, and implications for validity, are examined.

Building on research and experience with UDL, the third section discusses key categories of construct-irrelevant KSAs that can hamper students' learning and performance in assessments. It notes strategies to circumvent them, provide support, or mitigate their effects.

An integration of the ECD and UDL frameworks is then proposed. A support tool for test developers, called a design pattern, which integrates validity principles with UDL principles, is described (Haertel, DeBarger, Villalba, Hamel, & Colker, 2010). The design pattern provides support for the matching strategies described in Hansen, Mislevy, and Steinberg (2007), Hansen, Mislevy, Steinberg, Lee, and Forer (2005), and Kopriva (2008). The ideas are illustrated with examples from the Principled Science Assessment Designs for Students with Disabilities project (Haertel et al., 2010), supported by the Institute of Educational Sciences, U.S. Department of Education.

A more technical section then discusses psychometric foundations for conditional inference, using von Davier's (2008) General Diagnostic Model (GDM).[1] This psychometric framework is used to describe and compare the logic of four paradigmatic assessment situations:

- Marginal inference when the testing population is homogeneous with respect to having all the necessary construct-irrelevant KSAs the tasks require.

- Marginal inference when needed accommodations have not been used and the resulting mismatches are unknown to the score user.
- Conditional inference when task features and student construct-irrelevant capabilities are ascertained after testing occurs.
- Conditional inference when tasks are matched to students a priori.

The article focuses on assessment of special populations, but the principles can be applied more broadly. The closing section is a more general discussion of the theoretical and practical advantages of the approach.

## Assessment arguments

ECD is a framework that makes explicit, and provides tools for, building assessment arguments and assessments around the arguments (Mislevy & Riconscente, 2006; Mislevy, Steinberg, & Almond, 2003). ECD casts assessment as an argument from imperfect evidence. It aims to make explicit the claims, or the inferences one intends to make based on scores, and the nature of the evidence that supports those claims. It distinguishes layers at which different kinds of activities and structures appear in assessment design and operation, and provides tools and representations to support work at various layers.

This section describes ECD enough to coordinate the ideas that are central to the article: the roles of construct-relevant and irrelevant KSAs in validity, their relation to design choices about task features, UDL-infused design patterns that support task designers, and the connection to psychometric models.

### *ECD layers*

ECD sees the design process as first crafting an assessment argument, then embodying it in tasks, rubrics, scores, and procedures. Messick (1994) gives the essence of an assessment argument: "A construct-centered approach would begin by asking what complex of knowledge, skills, or other attribute should be assessed… Next, what behaviors or performances should reveal those constructs, and what tasks or situations should elicit those behaviors?" (p. 17).

The layers in ECD are Domain Analysis, Domain Modelling, the Conceptual Assessment Framework (CAF), Assessment Implementation, and Assessment Delivery (Table 1). They address, respectively, the substantive domain; the assessment argument; the structure of assessment elements such as tasks, rubrics, and psychometric models; the implementation of these elements; and their functioning in an operational assessment.

This article centres on the Domain Modelling and CAF layers. Domain Modelling is where assessment arguments are constructed; we will analyse the way that tailoring task features to learners impacts validity. The CAF is where the corresponding psychometric modelling takes place and particular task features are linked to learners' needs.

### *The structure of assessment arguments*

Messick's (1994) quote is a good start, but we need more machinery to examine the interplay of task design choices and validity. Figure 1 shows Toulmin's (1958) general schema for reasoning from particular data (D) to claims (C). A warrant (W; usually multifaceted, with backing (B) in research or experience) justifies this inference. In practice, we reason inductively, back up through the warrant, as indicated by the bold arrow from data to

Table 1. ECD layers.

| Layer | Focus of attention | Activities and Representations |
|---|---|---|
| Domain Analysis | The substantive domain | Determining what is important in the domain; i.e., what kinds of things do people need to know and do, in what kinds of situations. |
| Domain Modelling | The assessment argument | Arranging products of the Domain Analysis into the structure of assessment arguments. (Assessment arguments; Design Patterns) |
| Conceptual Assessment Framework | The structure of assessment elements | More formal & technical specifications for the elements of operational assessments. (Student, Evidence, and Task Models) |
| Assessment Implementation | Implementing the elements | Task and test assembly, fitting psychometric models, tuning scoring procedures. |
| Assessment Delivery | The functioning of the elements in an operational assessment | Architecture for assessment delivery (Almond, Steinberg, & Mislevy, 2002). |

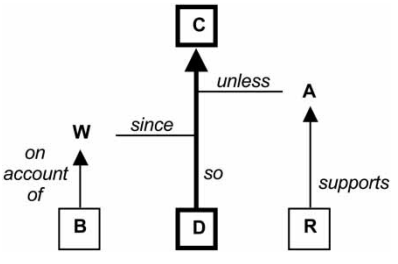


Figure 1. Toulmin's argument structure.

claim. Because data are rarely perfect or conclusive, we must usually qualify an inference in light of alternative explanations (A), which further "rebuttal" data (R) might support or weaken. Alternative explanations play a central role in validity and tailored task design.

Figure 2 applies the argument structure to assessment (Mislevy, 2006). We will focus on a single task, where "task" could range from a familiar multiple-choice item or an essay question to a language-proficiency interview or an open-ended problem in a computerized simulation. The claim is what we would like to say about some aspect of what a learner knows or can do. At the bottom of the diagram is a student's action in a situation: The student says, does, or makes something. The data is not the action itself, but our interpretations of the action and situation. There are three kinds of data:

- aspects of the person's actions;
- aspects of the situation; and
- additional information about the person's history or relationship to the observational situation.

The first of these is usually thought of as "the data" in assessment, but the second, the features of the situation, are equally necessary. The task must have features that engage the KSAs we are interested in. And if other task features present irrelevant impediments to a

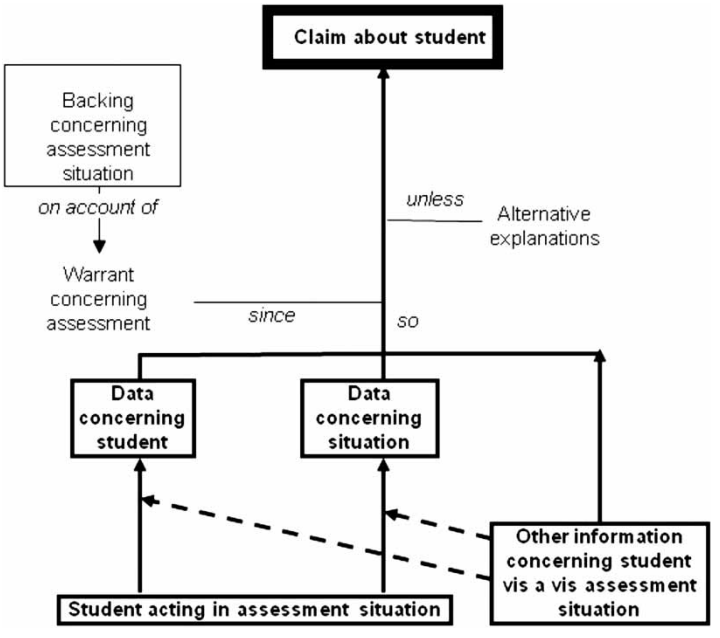


Figure 2. Extended Toulmin diagram for an assessment argument.

student's performance, we may not get meaningful evidence about what we are interested in. The third kind of data – what we know about the student with respect to the construct-irrelevant KSAs needed to access, interact with, and respond to tasks – helps us design tasks to minimize these obstacles.

A warrant about the targeted, or construct-relevant, KSAs comes with assumptions about access, interaction, and response capabilities. A warrant in genetics test might be,

> If a student understands how to form an inheritance-mode model to account for the coat colors of mice resulting from a crossing of two parents, then she will probably be able to fill in the cells of a Punnett square with the revised model.

The features of a corresponding task might include a diagram and text about a crossing and a computer tool to drag and drop genetics symbols into the Punnett square. A student might understand the required genetics, but perform poorly on the task because she is unfamiliar with the interface, or cannot distinguish colours that are associated with different alleles, or cannot physically manipulate the drag-and-drop device, or does not read English well enough to know what to do. All of these are alternative explanations for poor performance on the standard form of the task, other than the claim that she does not understand the genetics. The knowledge, skills, and abilities represented in the alternative explanations are construct irrelevant. A student needs them to succeed on the task, but they are not what the task is meant to assess. Presenting the identical task to all students derails those who are unable to perform due to construct-irrelevant KSAs, whether or not they have the genetics proficiencies.

### *Validity*

To get good evidence about the KSAs we care about, then, there are two things we need to do. First, we need to make sure that a task has features that are likely to elicit the targeted

KSAs, to the extent a student has them. A simulation task meant to probe understanding of Newton's laws that can be solved by trial and error gives "false positive" misleading information. Second, we need to make sure that the task does not require undue knowledge or skills that are unrelated to the targeted KSAs. A student who can work with Newton's laws but cannot figure out the simulation tool gives "false negative" misleading information. In both cases, alternative explanations are at play. Messick (1989) calls these threats to validity "construct under-representation" and "construct irrelevant sources of variance". They hold critical implications for task design.

### *Assessment design patterns*

A design pattern addresses a recurring design problem and the core of its solution. Alexander, Ishikawa, and Silverstein (1977) introduced design patterns in architecture, and they have been widely adapted in software engineering (Gamma, Helm, Johnson, & Vlissides, 1994). They capture experience and provide guidance, in a structure that builds in deep principles of a domain. They can be applied to resolve a problem in many situations even though the particulars are never exactly the same.

The Principled Assessment Design for Inquiry (PADI) project developed design patterns for assessment to provide a practical yet theory-based approach to develop high-quality assessments of science inquiry (Mislevy, Chudowsky, et al., 2003). PADI assessment design patterns conceptualize the elements of assessment arguments and their interrelationships as they apply to some targeted aspect of learning. They bridge the knowledge of content experts and measurement experts who need to work together to design complex assessment tasks. They help designers think through the mechanics of tasks in a way that leads to a coherent assessment argument (Table 2).

Each design pattern details three essential elements around which all assessments revolve: the student's knowledge, skills, and abilities about which one wants to make

Table 2. Attributes of a PADI assessment design pattern.

| Attribute | Definition |
|---|---|
| Title | Short name for the design pattern |
| Summary | Overview of the kinds of assessment situations students encounter in tasks supported by this design pattern. |
| Rationale | How tasks this design pattern supports provide evidence about the Focal KSAs. |
| Focal KSAs | Primary knowledge/skill/abilities of students that one wants to know about. |
| Additional KSAs | Other KSAs that may be required in tasks. |
| Potential observations | Features of performance that would provide evidence about the KSAs. |
| Potential work products | Different modes or formats in which students might produce the evidence. |
| Potential rubrics | Scoring rubrics that might be useful. |
| Characteristic features | Features of situations that are likely to evoke the desired evidence. |
| Variable features | Kinds of task features that can be varied in order to shift the difficulty or focus of tasks, or that require, circumvent, or support particular Additional KSAs. |
| Educational standards | Links to the related national, state, or professional standards. |
| Exemplar tasks | Links to sample assessment tasks that are instances of this design pattern. |
| References | Pointers to research and other literature that illustrate or provide backing for the design pattern. |

claims (*Focal KSAs*); the salient characteristics of what students say, do, or make that would provide evidence about acquisition of the Focal KSAs (*Potential observations*); and features of task environment that are needed to evoke the desired evidence (*Characteristic features*). (The last two concern data both as to what we would want to see students do and the situations they act in.)

*Rationale* articulates the warrant that justifies the targeted inferences and the kinds of task and evidence that support them. *Additional KSAs* may be required in a task that addresses the focal KSAs. Since Additional KSAs are not what is intended to be assessed, they can introduce threats to test validity. They need to be identified and minimized or avoided in order not to introduce construct-irrelevant variance. Alternatively, if it is known that the examinees possess a sufficient level of a given Additional KSA, that Additional KSA can be incorporated in the assessment tasks along with the intended KSAs. *Variable features* of tasks enable task developers to adjust the difficulty of tasks or focus their evidentiary value on different aspects of the Focal KSA, or to incorporate or circumvent particular additional KSAs.

*Potential work products* are students' responses or performances that are the source of Observations; different work products might require different combinations of Additional KSAs to produce. *Potential rubrics* are links to rules and instructions for evaluating work products. A design pattern also links to standards, other design patterns, task templates, and sample tasks to illustrate the connections in the design pattern.

The next section briefly reviews UDL. The section following that shows how design patterns can be used to integrate UDL with assessment arguments and validity.

## Universal design principles

The dialogue around student assessment now encompasses *all* students, as compared to past practices that excluded students with disabilities from accountability metrics. Beginning with the No Child Left Behind Act in 2001, states in the United States must include students with disabilities in reports of performance and progress. Developing assessment design frameworks that can produce tasks that are appropriate and accessible for a wide range of students requires new tools and approaches, including ones that can interface with frameworks for instructional and assessment materials (i.e., UDL) that are specifically designed to meet the needs of students with disabilities.

### *Rationale*

UDL helps to meet the challenge of diversity by suggesting flexible assessment materials, techniques, and strategies (Dolan, Rose, Burling, Harris, & Way, 2007). The flexibility of UDL empowers assessors to meet the varied needs of students and to accurately measure student progress. The UDL framework includes three overarching principles that address three critical aspects of any learning activity, including its assessment. *Multiple means of representation* addresses the ways in which information is presented. *Multiple means of action and expression* focuses on the ways in which students can interact with content and express what they are learning. *Multiple means of engagement* addresses the ways in which students are engaged in learning (Rose & Meyer, 2002, 2006; Rose et al., 2005).

#### *Principle I: provide multiple means of representation (the "what" of learning)*

Students differ in the ways they perceive and comprehend information that is presented to them. For example, those with sensory disabilities (e.g., blindness or deafness), learning

disabilities (e.g., dyslexia), and language or cultural differences may all require different ways of approaching content. Some may grasp information best when presented visually or through auditory means rather than printed text alone. Others may benefit from multiple representations of the content – a print passage presented with illustrative photographs or line drawings and an audio recording of the print passage.

*Principle II: provide multiple means of action and expression (the "how" of learning)*

Students differ in the ways they can interact with materials and express what they know. For example, individuals with significant motor disabilities (e.g., cerebral palsy), those who struggle with strategic and organizational abilities (executive function disorders, ADHD), and those who have language barriers, approach learning tasks differently, and will demonstrate their mastery differently. Some may be able to express themselves well in text but not in speech, and vice versa.

*Principle III: provide multiple means of engagement (the "why" of learning)*

Affect represents a crucial component to learning. Students differ in the ways in which they can be engaged or motivated to learn. Some enjoy spontaneity and novelty, while others prefer strict routine. Some persist with challenging tasks, while others give up quickly.

There is no one means of representation, expression, or engagement that will be optimal for all students in all assessment situations; providing multiple options for students is essential (CAST, 2012).

### *Categories of UDL*

In addition to the three principles that provide general guidance for infusing UDL into assessment design, we identify particular categories of student abilities (perceptual, expressive, language and symbols, cognitive, executive functioning, and affective) that are required for successful performance on assessment tasks but are usually not the assessment target. We want to use task features that support students who lack such construct-irrelevant abilities, or select features that minimize demand for them.

## Integrating ECD and UDL

A design pattern that integrates the principles of ECD and UDL can be used to create different versions of the kind of item that often appears on statewide science assessments, targeting the same construct-relevant KSAs but varying with respect to construct-irrelevant KSAs. An item called "Bicycle Rider"[2] illustrates the points.

### *Version A of Bicycle Rider*

The multiple-choice Bicycle Rider item was designed to test both middle-school science content and inquiry practices. The content is forces and motion. The inquiry practice is the ability to use appropriate tools and technologies to gather, analyse, and interpret data. The item describes how a person rides a bike at changing or constant speeds over time. It then asks the student to indicate which of four graphs, each illustrating a different relationship between speed and time, best characterizes the bicycle rider's travel.

Version A of the item presents students with a stem that reads "Which graph **best** represents the motion of a cyclist speeding up and then continuing at a constant speed?"

Under the stem, an image of four small graphs depicting possible speed (y-axis) over time (x-axis) relationships is presented, each labelled with a letter in the upper left corner. A link to enlarge the image in a separate window is provided underneath. Under the text and images, four radio buttons, labelled simply A through D, appear in a vertical array. A student is to choose the button that matches the letter of the graph she thinks is the best answer.

Version A appeared on a practice test from one state's large-scale middle-school science assessment. This assessment was developed and delivered by CAL Testing. Many features of the online testing platform were developed with UDL concerns in mind and included the following (Shaftel, Yang, Glasnapp, & Poggio, 2005):

- Progress monitoring on the screen ("breadcrumbs" across top of screen).
- Variable font size, magnifier, contrast.
- Text to speech.
- Radio buttons for multiple choice response capture.
- Testing environment tools: highlighter, striker, eraser, ruler, calculator.

The point is not that either version is UDL infused or not, but an examination through the UDL/validity perspective of two variants from a potential family of tasks that tap the same construct but with features that require different combinations of Additional KSAs.

### *A UDL-infused design pattern*

Version A of Bicycle Rider is aligned with a design pattern entitled "Interpreting Data in Tables, Charts, and Graphs". This design pattern was developed in collaboration with a state department of education in the *Principled Science Assessment Design for Students with Disabilities* project (Haertel et al., 2010; Rose et al., 2012; Zhang et al., 2010). The complete design pattern can be accessed at http://design-se.padi.sri.com/padi/AddNodeAction.do?NODE_ID = 2140&state = viewNode. This design pattern supports creating items that involve understanding and interpreting data and data-variable relationships as represented in tables, charts, or graphic forms. Since every science content area involves data, it supports item development in all areas.

This design pattern also integrates principles of universal design for learning (UDL) into specific design pattern attributes. Haertel et al. (2010) provide a more detailed discussion, but the key ideas are these:

- Focal KSAs are knowledge, skills, or other attributes that are the focus of the design pattern, and are usually construct relevant in a task the design pattern supports. They are intimately connected with the characteristic task features. For example, the focal KSAs to be assessed in Bicycle Rider are:
  - ability to compare and /or contrast multiple representations and the data represented therein;
  - ability to describe simple mathematical relationships or trends among data;
  - ability to draw conclusions or make predictions based on data.
- Additional KSAs identify other knowledge and skills that may or may not be construct relevant but are required to successfully answer the item; the assessment designer determines whether the Additional KSAs will be supported through the use of variable features in the assessment or whether they will remain unaddressed.

Examples of two Additional KSAs that are prerequisite knowledge for successfully completing an item like Bicycle Rider are:

- awareness of different representational forms;
- knowledge of what data are.

Additional KSAs that represent learner needs (UDL) for successful performance on an item also appear in Additional KSAs, organized in the six categories listed in the UDL section. We will see examples when we compare the two versions of the item.

- Characteristic features of tasks must be present if the task is to provide evidence about the construct. Making sure the task embodies all of the characteristic features ensures that the construct-relevant KSAs will be assessed. For items like Bicycle Rider, characteristic features include:
  - the presentation contains numeric data;
  - the presentation includes at least one representational form;
  - the presented data are in a scientific context.
- Variable task features include ones that allow a designer to adjust the difficulty, the scope, and the focus of a task while obtaining evidence about the construct. UDL-infused design patterns, in particular, detail features that can be varied to support, bypass, or appropriately target demands for construct-irrelevant KSAs. They are linked to the relevant Additional KSAs, so that clicking on a particular Additional KSA highlights variable task features that can be manipulated to increase, decease, or support demands for that KSA. Variable features a designer can consider when writing an item like Bicycle Rider are given below. The next section shows how some of these Variable Features were used to change demands for particular Additional KSAs in Version B of Bicycle Rider.
  - Number of representations.
  - Complexity of representational form(s).
  - Number of variables represented in the table, graph, or chart.
  - Perceptual features: visual information (e.g., automatic text to speech).
  - Language and symbols: supports for syntactic skills and underlying structure (e.g., highlighted syntactical elements).
  - Cognitive features: supports for critical features, big ideas, and relationships (e.g., examples to emphasize critical concepts).
  - Executive features: supports for managing information (e.g., locate radio buttons near relevant images).
  - Affect features: supports for intrinsic motivation (e.g., enhanced relevance with real world context).
- Potential observations are suggested to gather evidence about the focal KSAs. For example, if a task is intended to assess the focal KSA "the ability to compare and contrast multiple representations and the data represented therein", then one potential observation would involve having the student read a description of a data relationship and select from among several graphical representations the one that most accurately depicts it. In the design pattern, the relevant potential observation is:
  - identification of representational forms of data that communicate the same mathematical relationships among data (or trends in data).

- Potential work products indicate a form in which students can produce responses. Work products can vary in ways that are sensitive to resource constraints and logistical considerations. In UDL-infused design patterns, students with varying profiles of needs (e.g., visually impaired, limited dexterity) may require different work product options. The entries in Potential Work Product and Additional KSA are linked to help task designers see the connections. An example of a work product is:

    - selection of a representation to match a data relationship.

***Version B of Bicycle Rider***

Both versions of Bicycle Rider address UDL concerns at the level of the testing platform. We introduced variations motivated by the design pattern in the specifics of Version B. Possible sources of construct-irrelevant variance were identified related to individual students' learning needs in terms of perception, expression, language and symbols, cognition, executive functioning, and engagement (affective). These categories of needs appear as Additional KSAs in the design pattern. These Additional KSAs are linked to variable task features in the design pattern that could be used to support students' needs for construct-irrelevant KSAs. These variable features were used to identify revisions that would reduce particular construct-irrelevant demands, while maintaining the characteristic features and thus the construct-relevant demands. These UDL-motivated variations were thus constrained by the essential focus of the assessment argument.

The item stem of Version B reads (with bolding): "A student rides her bike to school. She first **speeds up for 1 minute.** She then **continues at a constant speed for 5 minutes.** [line break.] Which graph best represents her motion over this time?" Under the stem, four time-by-speed graph answer options appear. They show the same relationships as the graphs in Version A, but they are visually improved through enlargement and clearer labelling. Furthermore, a radio button appears directly to the top left of each y axis, obviating a letter-based choice, and thus eliminating a set of correspondences to identify and the associated demand for working memory. For students above these hurdles, the variant would differ negligibly; its difficulty would be driven by their understanding of the graph-motion representation. For students challenged by the construct-irrelevant KSAs, Variant B poses less demand for these KSAs. They might still answer incorrectly if they do not understand the representation, but the same response would provide more valid information since the alteration reduces the force of alternative explanations for missing the item because of demands for construct-irrelevant KSAs.

Note that this logic reveals a shortcoming of using "differential boost" (Fuchs, Fuchs, Eaton, Hamlett, & Karns, 2000) to evaluate the effectiveness of UDL-motivated item revisions. It can be that a revision reduces a need for a construct-irrelevant KSA for many students, and for more students *without* disabilities who have the focal skill than students *with* disabilities who have the skill. The group without disabilities now has even fewer false-negative wrong answers. Validity has increased, yet the pattern in percents-correct is the opposite of differential boost.

Table 3 summarizes the UDL variations implemented in Version B. The Interpreting Data in Tables, Charts, and Graphs design pattern (http://design-se.padi.sri.com/padi/AddNodeAction.do?NODE_ID = 2140&state = viewNode) shows how the extended UDL features are represented in the design pattern template.

Two examples of principled changes between the Version A and B items illustrate the UDL design logic. First, note that in the stem in Version A, no context is given for the ride

Table 3. UDL principles (categories of students' needs) supported by variable features in Version B of Bicycle Rider

| UDL Principle (Category of Student Need) | Task Model Variables Implemented to Address UDL Principles in Version B of Bicycle Rider |
|---|---|
| Perceptual Features | -Flexible size of text and images<br>-Flexible amplitude of speech and sound<br>-Adjustable contrast<br>-Flexible layout<br>-Visual graphics<br>-Verbal descriptors (spoken equivalents for text and images)<br>-Automatic text to speech |
| Skill and Fluency | -Alternative to written response (radio buttons) |
| Language and Symbols | -Embedded support for key terms,<br>-Alternate syntactic levels (simplified text)<br>-Support for decoding (digital text and automatic text to speed) |
| Cognitive Features | -Using explicit examples to emphasize critical concept (minutes cyclist accelerating and at constant speed)<br>-Presentation of graphical representation simultaneously as compared to one at a time (reduce cognitive load) |
| Executive Features | -Reduced working memory<br>-Locate radio buttons near relevant images on-screen<br>-Progress monitoring |
| Affect Features | -Real-world context to heighten engagement<br>-Age-appropriate materials |

or the amount of time it takes. The stem in Version B says who is riding the bike and the amount of time the ride takes. This second version was guided by the Cognitive and Affective UDL categories, providing a real-world context and explicit designation of time. There is a tradeoff: Version A has fewer words and does not introduce a potentially distracting context, whereas Version B may increase engagement. There will be some students for whom Version A better matches their profile of construct-irrelevant KSAs, and other students for whom Version B does.

Second, the graphs in Version A are assigned a letter (A through D) that must be referenced in the array underneath to make the radio-button answer choice. In Version B, the radio buttons appear directly adjacent to the graphs, eliminating the letter choice translation. Minimizing steps speaks directly to the Skill and Fluency, Cognitive, and Executive Functioning UDL categories. These UDL-motivated variations of task features were facilitated by the technology platform that supports both versions, illustrating how computer delivery can produce tailored forms from an item family to individual students.

## A psychometric framework

The previous sections showed how the UDL and ECD frameworks could be integrated, so one can construct tasks that evoke comparable evidence even though surface features vary to tap combinations of construct-irrelevant KSAs that match the capabilities of different students. This section provides a psychometric framework for inference in an assessment designed according to these principles. It expresses the argument structures can be expressed in terms of von Davier's (2008) general diagnostic model (GDM). The GDM is a member of a class of models called cognitive diagnosis models

(Leighton & Gierl, 2007) or diagnostic classification models (Rupp, Templin, & Henson, 2010).

### *Key ideas*

A cognitive diagnosis model gives probabilities of task responses as functions of features of tasks and student proficiencies associated with those features. The analyst indicates which "attributes" are involved in each task; students are similarly characterized in terms of their proficiency with respect to the same set of attributes.

The simplest cognitive diagnostic models have dichotomous attributes: Tasks do or do not require them, students do or do not have them. For example, Tatsuoka's (1983) mixed-number subtraction example characterized students in terms of which algebraic procedures a student has mastered, and tasks in terms of which procedures they require. Probabilities of success are modelled in terms of how a student's profile of proficiencies on the attributes matches up with the profile of attributes an item requires. The ideas extend to more complicated response variables, such as counts and ordered category responses, and to more complicated attributes of people, such as ordered categorical states (such as level on a learning progression) and continuous variables (such as decoding skill).

Cognitive diagnosis models allow a range of ways to combine student proficiencies to model item response probabilities. They support compensatory combinations (being high in some proficiencies can make up for being low in others), and combinations such disjunctions, when different proficiencies can be employed to succeed on a task, and conjunctions, for when certain proficiencies are necessary jointly. We can use conjunctive combinations to model the effect of necessary but construct-irrelevant KSAs. We can posit, for example, that a task requires a conjunctive combination of several construct-irrelevant KSAs and a set of construct-relevant KSAs, which are effective only given a sufficient value on the conjunction of construct-irrelevant KSAs.

Taking advantage of these combination properties, we will lay out a quantitative model for the logic of four testing situations that can also be described as qualitative assessment argument structures. We use the vector of attribute values to describe which construct-irrelevant and construct-relevant demands are designed into a task variant. We use the vector of attribute variables to describe students' profiles of construct-irrelevant and construct-relevant KSAs. We use the structure of the model to express what is likely to happen when a particular student is presented a particular variant of a task. We show how this conceptualization fits the strategy of administering task variants for each student to give them the best chance to show what they know and can do.

### *A general diagnostic model*

The key elements of a cognitive diagnosis model are contained in the general form $p(\mathbf{X} = \mathbf{x}|\boldsymbol{\theta}, \mathbf{Q}, \boldsymbol{\eta})$, where $\mathbf{X} = (X_1, ..., X_n)$ represents $n$ task response variables and $\mathbf{x} = (x_1, ..., x_n)$ values they can take; $\boldsymbol{\theta} = (\theta_1, ..., \theta_K)$ is a vector of $K$ proficiency variables (KSAs) that categorize a student ("attributes" in cognitive diagnosis terminology); $\mathbf{Q}$ is a matrix with $n$ rows, one per task, with the $j^{\text{th}}$ row a vector $\mathbf{q}_j = (q_{j1}, ..., q_{jK})$ indicating the qualitative association of Task $j$ to the $K$ attributes; and $\boldsymbol{\eta}$ is a vector of parameters that gives the quantitative relationship between task features and students' probabilities of success. This expression indicates that there are multiple aspects of students' knowledge and skill; that tasks have features we can relate to these proficiencies in known ways, by

virtue of the tasks' construction; and in a manner to be specified, how they interact cognitively determines how students are likely to respond.

Particular forms of these variables and functions must be specified. The following basic specifications allow us to make observations about conditional inference:

- Assume all items are dichotomous, where $x_j = 1$ indicates a correct response and 0 indicates incorrect.
- Partition $\boldsymbol{\theta}$ into $(\phi_1, ..., \phi_K, \theta)$, where the $\phi_k$s are construct-irrelevant KSAs and $\theta$ is the construct that is the target of measurement. Eliciting evidence about $\theta$ necessarily requires some construct-irrelevant KSAs to access, interact with, and respond to a task.
- Define the task-attribute vectors $\mathbf{q}_j$ such that $q_{jk}$ indicates the demand for construct-irrelevant KSA $k$ required by Task $j$. Some of the elements of $\mathbf{q}_j$ may also be defined in terms of the presence of supports or accommodations, as might be found in students' Individualized Education Programs (IEPs). In this case, the corresponding $\phi$s indicate a student's need for such supports or accommodations. *All* tasks are constructed to have some level of demand on $\theta$. They can differ as to which $\phi_k$s they require and what amounts. We can define families of tasks that are equivalent as to their construct-relevant demands, but differ as to construct-irrelevant demands (Kopriva, 2008).
- Define the combination functions $h_k(q_{jk}, \phi_k)$ to take the value 1 if a student's value of $\phi_k$ equals or exceeds the level of demand for KSA $k$ that is required in Task $j$, and 0 if not. That is, $h_k(q_{jk}, \phi_k) = 1$ means that the student is *above the hurdle* with respect to the demands for KSA $\phi_k$ posed by Task $j$; for example, whether a student's visual acuity makes it possible for her to read the font-size of Task $j$. If a task has no demand for a $\phi_k$, then $h_k(q_{jk}, \phi_k) = 1$. When an element of $\mathbf{q}_j$ is defined in terms of the presence of a support or accommodation, $h(q_{jk}, \phi_k) = 1$ if either the student does not need the support of accommodation or if she needs it and it is present. Here $h(q_{jk}, \phi_k) = 0$ means the student needs the support or accommodation but Task $j$ does not provide it. Again $h_k(q_{jk}, \phi_k) = 1$ means the student is *above the hurdle*.
- Define the construct-relevant combination function $f(\theta,\beta_j)$ as a standard psychometric model, such as an item response theory (IRT) model in which the probability of a correct response is a function of a student's $\theta$ and characteristics of Task $j$ such as its difficulty with respect to $\theta$.
- Let $\pi_j$ be the (chance) probability of a student getting Task $j$ right even if he is not above the hurdle on one or more construct-irrelevant KSAs, that is, $\phi_k$s. (Together, the $\pi_j$ s and $\phi_k$s constitute $\boldsymbol{\eta}$ in the GDM.)

The form of the probability model that accommodates conditional-inference is then

$$\Pr\left(x_j = 1 \middle| \phi_1, ..., \phi_K, \theta, \mathbf{q}_j, \beta_j, \pi_j\right) = \pi_j + (1 - \pi_j) \prod_k \left[h_k(\phi_k, q_{jk})\right] f(\theta, \beta_j). \quad (1)$$

By the way that the $q_{jk}$s, $\phi_k$s, and $h(q_{jk}, \phi_k)$s are defined)s are defined, $\prod \left[h_k(\phi_k, q_{jk})\right] = 0$ if there is at least one $k$ for which Task $j$'s demand with respect to construct-irrelevant KSA $\phi_k$ exceeds the student's capabilities. In this case, the entire second addend is 0; the probability of getting the item right is just $\pi_j$, and the response does not depend on $\theta$ at all. If, on the other hand, for every $\phi_k$ there is either no demand or the demand is within the student's capabilities (i.e., she is "above the hurdle" for those construct-irrelevant

KSAs), then $\prod \left[ h_k\left(\phi_k, q_{jk}\right) \right] = 1$ and the probability of a correct response depends on $\theta$. It may be a high probability or a low probability, but it depends on the targeted KSAs. This is a mathematical expression to say that valid inference about the construct is *conditional* on the task's necessary construct-irrelevant KSA demands being satisfied.

This model can be extended in many ways, such as alternative response types and multivariate $\theta$s. Another extension is more gradual $h$ functions. Instead of all or nothing, over the hurdle or not, one can model performance that gradually degrades as a student falls farther below a task's demand for some $\phi_k$s.

Putting these ideas into practice requires specifying the structure of **Q** and the forms of the $h_s$s and $f$. Strategies and tools for doing so are appearing in the cognitive diagnosis literature (e.g., Kunina-Habenicht, Rupp, & Wilhelm, 2012; Rupp et al., 2010). We note that this research concludes that it is much better to start with strong hypotheses from theory and experience, to build tasks and accommodation options around these frameworks, then fine-tune specifications, rather than write tasks and try to come up with $\phi$s and **Q**s post hoc.

The conditional framework introduces a responsibility to test designers and test users to demonstrate that alternative, not-surface-equivalent, forms of tasks do in fact provide equivalent evidence about students. Some instances will not be straightforward. When students can choose items, for example, they may make choices that disadvantage them (Wainer & Thissen, 1994). When variants differ in terms of the language in which they are presented, literal translation need not result in equivalence with respect to construct-relevant demands; more thoughtful adaption with respect to cultural as well as linguistic matters is required (Hambleton, Merenda, & Spielberger, 2004). Experiments with different forms of tasks can shed light on the interacting demands of construct-relevant and -irrelevant task features (Abedi, Lord, Hofstetter, & Baker, 2005). Design strategies and analytical tools developed in these specific areas can be adapted to implementation of the conditional assessment paradigm more generally.

### *Four inferential situations*

We can use the GDM for conditional inference to examine four paradigmatic assessment situations, under the assumption that Equation (1) is the correct model.

#### *Marginal inference when all students are above all construct-irrelevant KSA hurdles*

The traditional standardized testing situation, before the introduction of supports or accommodations, assumed a homogeneous population, in the following sense: All students were assumed to have sufficient capabilities in all construct-irrelevant KSAs required by all the items in the test. If this were so, $\prod \left[ h_k\left(\phi_k, q_{jk}\right) \right] = 1$ for all students, their performance is direct evidence about $\theta$ through $f(\theta,\beta_j)$; and, since the everyone-over-all-$\phi$ hurdles is correct, it is not necessary to include the $\phi$s and the **q**s in the operational model. All of the systematic variation among students' performances is assumed to be due to variation in their $\theta$s. When this case holds, familiar scores, whether through classical test theory or IRT, support valid inferences about the construct. Under these conditions, equivalent surface conditions do indeed help provide equivalent evidence about students.

#### *Marginal inference when, unbeknown to the score user, some students are not above construct-irrelevant hurdles*

This is the case we want to avoid: All students are administered the standard form of the test, with its items' varying **q** features and their consequent $\phi$ demands, and some students are not above all the hurdles on all the items. That is, there are at least some items and

students for which $h(q_{jk}, \phi_k) = 0$. If tests are scored in the usual way, assuming that the preceding case holds, then inferences about students' $\theta$s are obtained through $f(\theta,\beta_j)$ for any such item. The student's performance on such items, however, is spuriously low, at $\pi_j$, and because $\prod \left[h_k(\phi_k, q_{jk})\right] = 0$, the response contains no information about her $\theta$.

Even if a student is above all the hurdles, so $\prod \left[h_k(\phi_k, q_{jk})\right] = 1$, she still might not have a high probability of answering correctly if her $\theta$ is low. This is what we want scores to tell us. Getting an item wrong due to lack of $\phi$ and getting it wrong due to a low $\theta$ have identical observed data, an incorrect response. It is misleading as evidence about $\theta$ in the first case, while it is relevant in the second.

An interesting situation arises when we do not know whether students are above the hurdles for construct-irrelevant KSAs, but it just so happens that they all are. Their responses depend on only $\theta$. Had we ensured beforehand that all were above the hurdle, we would see exactly the same data, analyse it with the same model $f(\theta,\beta_j)$, and get exactly the same scores – but inferences from them would be more valid, because we rule out alternative explanations for poor performance due to construct-irrelevant demands.

*Conditional inference when task features and student construct-irrelevant capabilities are inferred after testing occurs*

Here students are tested with surface-equivalent forms, but we use a full model like Equation (1). It is sometimes possible, using the statistical machinery of cognitive diagnosis, to infer students' patterns of $\phi$s from the patterns of their responses (Rupp et al., 2010). Doing so requires careful construction of items and tests so the **q**s are known and properly balanced. It is then possible to obtain inferences about students' $\theta$s from their response patterns, carrying out *conditional inference by analysis*. The assessment situation is the same as the second case described above, but now we are using an appropriate psychometric model to try to sort out the influence of construct-relevant and construct-irrelevant demands.

This approach is usually unsatisfying, because the added uncertainty that comes from trying to estimate the $\phi$s jointly with $\theta$s renders the estimates unreliable. The lessons for this case are that (a) it may be possible to carry out valid conditional inference using an appropriate model, (b) the appropriate model is not the standard model for marginal inference, and (c) doing this is generally not practical.

*Conditional inference when tasks are matched to students a priori*

This situation obtains when (a) students vary meaningfully with respect to the construct-irrelevant KSAs that are necessary to access, interact with, or respond to assessment tasks; (b) we know how they vary, such as by having their IEPs or knowing what prerequisite knowledge they have; (c) we have or can construct items such that their demands for construct-irrelevant KSAs are known; and (d) we assign to each student, for each item, a variant for which $\prod \left[h_k(\phi_k, q_{jk})\right] = 1$. Now scores depend on $\theta$, not $\phi$s. Inference about $\theta$ from the assessment is more valid, because we have weakened alternative explanations for poor performance due to lack of necessary $\phi$s. Because we have done the required work in matching students with task variants, we can again use the simple test scoring models $f(\theta,\beta_j)$ and carry out *conditional inference by design*. Compared to the previous case, the modelling demands are lower and the reliability as well as the validity of scores is higher.

## Conclusion

Both theoretical and practical benefits accrue from integrating UDL and ECD in the assessment design process.

### *The central benefit of fairness*

Increasing fairness in assessment by integration of ECD and UDL principles from the start has been the goal of our work. The *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council of Measurement in Education, 1999) sees fairness as fundamental to test validity, and specifically addresses incorporating UDL to help develop tests that are fair to all examinees. Our goal to build "fair" assessments is expressed in thoughtfully applying the ideas of ECD to provide all students with an opportunity to perform at their best in assessment situations. Infusing UDL into the design process from the very beginning is a principled, proactive way to reduce accessibility barriers of assessment tasks, in contrast to retrofitting them individually.

Much of the practice of ECD is focused on the identification of sources of construct-irrelevant variance that can result in flawed inferences from scores. Unexamined design choices can lead to tasks that use unnecessarily unfamiliar language and syntax, poorly understood social and cultural item contexts and task stimuli, or modes of representations (visual, aural, behavioural) that may be systematically biased against subgroups with limited access to requisite background knowledge or use of sensory modalities. Fairness requires that task contexts be sufficiently familiar, appropriate, and accessible to all students. Articulation of task models from the beginning of the process reduces the incidence of tasks that disadvantage particular students.

### *Theoretical benefits*

There is a growing body of research and practical experience with assessments meant to serve more diverse student populations. Educative, moral, and legal imperatives motivate the work. Various projects investigate problems from perspectives of special education, educational technology, and domain learning, but there are gaps and conceptual mismatches across experts from different backgrounds. This article provides a unified theoretical framework for assessment design and analysis.

A unified framework helps bring together principles from different fields. A UDL-infused design pattern, for example, not only brings in insights from educational technology, science learning research, and special education, it does so in a form that makes the connections for test designers and helps them build tasks with valid assessment arguments from the outset.

The extension to a psychometric framework further aides practice, since there has been little connection between the psychometric community and the UDL community. The work articulates the vision of fairness arising from the UDL and special needs communities with the models of performance and formal statistical inference from the psychometric world. The framework makes it possible to take advantage in a principled way of the opportunities that computer-administered testing affords. Test delivery systems are now available that can adapt in real time to student needs and provide choices to students to support construct-irrelevant KSAs (Shaftel et al., 2005). Given prior information about students such as their IEPs, it becomes feasible to assemble instances of task models that vary in their surface characteristics but provide equivalent evidence about the targeted construct (Hansen et al., 2005).

### *Practical benefits*

#### *Increased engagement*

State-of-the-art assessment design now includes the use of context-rich, situated tasks often presented in online or computer-based environments. These tasks often involve story narratives to increase student engagement and motivation and present students with conceptual links previously unavailable in paper-and-pencil testing. Technology-enhanced tasks also support the use of open-ended, interactive contexts that focus on student reasoning processes, permit multiple solution paths, and present varied stimuli and concepts that were impossible in paper-pencil assessment.

The same characteristics of technology-enhanced tasks that help assess students' extended reasoning may present accessibility barriers to students with disabilities. Students with cognitive disabilities, for example, may be overwhelmed with extended reasoning tasks due to cognitive load and memory or executive functioning demands. Research has shown that some combinations of stimuli can overwhelm students' working memory. Chandler and Sweller (1992) documented a split attention effect where students' learning was hampered by the combination of animation, narration and on-screen text, as compared to just animation and narration.

An ECD process can guide designers in the application of UDL principles as they consider ways to recruit interest, sustain effort, and provide options for self-regulation. For example, designers might consider ways that students can monitor their progress as they work through a task. Variable task features that could help students monitor their progress include a progress bar, intermittent messages to the student about their progress, and interactive navigation through an extended task.

#### *Linking instructional practices to student needs*

In ECD's domain modelling layer, design elements reflect the assessment of that domain but also reflect aspects of learning. Designers specify Focal KSAs for assessment, which are also key learning goals (Krajcik, McNeill, & Reiser, 2008). Work products that provide evidence of proficiency in a domain also characterize the activities that students engage in for learning. Because the ECD process has required identifying Additional KSAs, students' learning needs have also been identified as well. The linkage among the additional KSAs and the variable features applies not only to assessment but to day-to-day instruction.

A student's performance on a prerequisite learning goal can also provide information about an Additional KSA, to be supported or built on accordingly by the classroom teacher. Additional KSAs thus provide teachers with information about whether a student has the knowledge needed to acquire the new learning target. Assessment tasks and associated instructional activities can be designed to support the cognitive load that students encounter in multi-step, complex learning goals and problem situations. Teachers can use the design patterns and the specification of additional KSAs associated with Focal KSAs to create instructional activities that support student's learning needs.

Variable task features articulated in design patterns thus correspond to the very same scaffolds that are the critical features of instruction, to ensure that instructional content is accessible to students. For example, using multiple representations in instruction can help make instructional concepts salient (Ainsworth, 2006), and can be used analogously in assessment design to ensure that target KSAs are presented in multiple ways and remain the primary focus of a task. Similarly, vocabulary support, demonstrations of skills, and contrasting cases can be used in both instruction and assessment. While the

present article focuses on large-scale assessment, these ideas apply equally to classroom testing.

## Acknowledgements

Research findings and assessment tasks described in this article were supported by the Principled Assessment Science Assessment Designs for Students with Disabilities (Institute of Education Sciences, US Department of Education, R324A070035). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies. We are grateful to the guest editor Hossein Karami, Heather Buzick, Eric Hansen, Shelby Haberman, and two anonymous reviewers for helpful suggestions, to John Poggio (University of Kansas), Richard Vineyard (Nevada State Department of Education), and Abel Leon (CAL Testing) for their generous support during the implementation of assessment tasks at schools and school districts, to Robert Dolan as co-PI at the start of the project and advisor thereafter, and to Eric Hansen for advice on evidence-centered design and task accommodations.

## Notes

1. Hansen et al. (2007) and Hansen et al. (2005) give an alternative specification using Bayes nets.
2. Released item in the "7th Grade Science Formative Test" by CAL Testing (formerly Kansas Computerized Assessments). Retrieved April, 2009 from http://kca.cete.us, Center for Educational Testing and Evaluation.

## References

Abedi, J., Lord, C., Hofstetter, C., & Baker, E. (2005). Impact of accommodation strategies on English language learners' test performance. *Educational Measurement: Issues and Practice*, *19*(3), 16–26.

Ainsworth, S. (2006). DeFT: A conceptual framework for considering learning with multiple representations. *Learning and Instruction*, *16*, 183–198.

Alexander, C., Ishikawa, S., & Silverstein, M. (1977). *A pattern language: Towns, buildings, construction*. New York, NY: Oxford University Press.

Almond, R. G., Steinberg, L. S., & Mislevy, R. J. (2002). Enhancing the design and delivery of assessment systems: A four-process architecture. *Journal of Technology, Learning, and Assessment*, *1*(5). Retrieved from http://www.bc.edu/research/intasc/jtla/journal/v1n5.shtml

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

CAST. (2012). *Universal design for learning guidelines Version 2.0*. Wakefield, MA: Author.

Chandler, P., & Sweller, J. (1992). The split-attention effect as a factor in the design of instruction. *British Journal of Educational Psychology*, *62*, 233–246.

Dolan, R. P., Rose, D. H., Burling, K., Harris, M., & Way, D. (2007, April). *The universal design for computer-based testing framework: A structure for developing guidelines for constructing innovative computer-administered tests.* Paper presented at the National Council on Measurement in Education Annual Meeting, Chicago, IL.

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*, 65–85.

Gamma, E., Helm, R., Johnson, R., & Vlissides, J. (1994). *Design patterns*. Reading, MA: Addison-Wesley.

Green, B. (1978). In defense of measurement. *American Psychologist*, *33*, 664–670.

Haertel, G., DeBarger, A. H., Villalba, S., Hamel, L., & Colker, A. M. (2010). *Integration of evidence-centered design and universal design principles using PADI, an online assessment design system* (Technical Report 3). Menlo Park, CA: SRI International.

Hambleton, R. K., Merenda, P. F., & Spielberger, C. D. (Eds.). (2004). *Adapting educational and psychological tests for cross-cultural assessment.* Mahwah, NJ: Erlbaum.

Hansen, E. G., Mislevy, R. J., & Steinberg, L. S. (2007). *Accessibility of testing within a validity framework* (U.S. Patent # 7217134). Washington, DC: U.S. Patent and Trademark Office.

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

Kopriva, R. J. (2008). *Improving testing for English language learners.* Philadelphia, PA: Psychology Press.

Krajcik, J., McNeill, K. L. & Reiser, B. J. (2008), Learning-goals-driven design model: Developing curriculum materials that align with national standards and incorporate project-based pedagogy. *Science Education*, *92*, 1–32.

Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2012). The impact of model misspecification on parameter estimation and item-fit assessment in log-linear diagnostic classification models. *Journal of Educational Measurement*, *49*, 59–81.

Leighton, J., & Gierl, M. (Eds). (2007). *Cognitive diagnostic assessment for education: Theory and applications*. New York, NY: Cambridge University Press.

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

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

Mislevy, R. J. (2006). Cognitive psychology and educational assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 257–305). Phoenix, AZ: Greenwood.

Mislevy, R. J., Chudowsky, N., Draney, K., Fried, R., Gaffney, T., Haertel, G., … Wilson, M. (2003). *Design patterns for assessing science inquiry* (PADI Technical Report 1). Menlo Park, CA: SRI International.

Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice*, *25*, 6–20.

Mislevy, R. J., & Riconscente, M. M. (2006). Evidence-centered assessment design: Layers, structures, and terminology. In S. Downing & T. Haladyna (Eds.), *Handbook of test development* (pp. 61–90). Mahwah, NJ: Erlbaum.

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

Rose, D. H., & Meyer, A. (2002). *Teaching every student in the digital age: Universal design for learning.* Alexandria, VA: ASCD.

Rose, D. H., & Meyer, A. (Eds.). (2006). *A practical reader in universal design for learning.* Cambridge, MA: Harvard Education Publishing Group.

Rose, D., Meyer, A., & Hitchcock, C. (Eds.). (2005). *The universally designed classroom*. Cambridge, MA: Harvard Education Press.

Rose, D., Murray, E., & Gravel, J. (2012). *UDL and the PADI process: The foundation* (Technical Report 4). Menlo Park, CA: SRI International.

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications.* New York, NY: Guilford Press.

Shaftel, J., Yang, X., Glasnapp, D., & Poggio, J. (2005). Improving assessment validity for students with disabilities in large-scale assessment programs. *Educational Assessment*, *10*, 357–375.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, *20*, 345–354.

Toulmin, S. (1958). *The use of argument.* Cambridge, UK: Cambridge University Press.

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*, *61*, 287–307.

Wainer, H., & Thissen, D. (1994). On examinee choice in educational testing. *Review of Educational Research*, *64*, 159–195.

Zhang, T., Mislevy, R. J., Haertel, G., Javitz, H., Murray, E., Gravel, J., & Hansen, E. G. (2010). *A design pattern for a spelling assessment for students with disabilities* (Technical Report 2). Menlo Park, CA: SRI International.