

Assessment for Students with Disabilities
Technical Report 5 | February 2013



Conditional Inferences Related to Focal and Additional Knowledge, Skills, and Abilities

Project: Principled Science Assessment Designs for Students with Disabilities

Robert J. Mislevy, Educational Testing Service

Geneva Haertel, Britte H. Cheng, Daisy Rutstein, and Terry Vendlinski, SRI International

Elizabeth Murray and David Rose, CAST

Jenna Gravel, Harvard University

Alexis Mitman Colker, Independent Consultant

Report Series Published by SRI International





SRI International
Center for Technology in Learning
333 Ravenswood Avenue
Menlo Park, CA 94025-3493
650.859.2000
<http://padi-se.sri.com>

Technical Report Series Editors

Alexis Mitman Colker, Ph.D., *Project Consultant*
Geneva D. Haertel, Ph.D., *Co-Principal Investigator*
Robert Mislevy, Ph.D., *Co-Principal Investigator*
Ron Fried, *Documentation Designer*

Copyright © 2013 SRI International. All Rights Reserved.

**Conditional Inferences Related to
Focal and Additional Knowledge, Skills, and Abilities**

Robert J. Mislevy, Educational Testing Service

Geneva Haertel, SRI International

Britte H. Cheng, SRI International

Elizabeth Murray, Center for Applied Specialized Technologies (CAST)

David Rose, Center for Applied Specialized Technologies (CAST)

Jenna Gravel, Harvard University

Alexis M. Colker, Independent Consultant

Daisy Rutstein, SRI International

Terry Vendlinski, SRI International

Abstract

Standardizing aspects of assessments has long been a tactic to help make fair evaluations of examinees. The idea is to reduce variation in irrelevant aspects of testing procedures that could advantage some examinees and disadvantage others. However, recent efforts to make assessment available to a more diverse population of students has highlighted 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 programs, for the most part these practices lie outside formal educational measurement theory. This paper builds on recent research in universal design for learning (UDL), assessment design, and psychometrics to explicate the rationale for inference that is conditional on matching examinees with principled variations of an assessment so as to minimize construct-irrelevant demands. Examples of the logic are illustrated using an item from a state large-scale science assessment.

Key words: evidence-centered design, general diagnostic model, universal design for learning.

Acknowledgements

Research findings and assessment tasks described in this paper 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 Prof. 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.

Validity, reliability, comparability, and fairness are not just measurement issues, but social values that have meaning and force outside of measurement wherever evaluative judgments and decisions are made.

Samuel Messick, 1994, p. 2.

1. Introduction

Traditional test formats standardize the materials and the circumstances of test procedures, epitomizing one sense of “fairness”: All examinees are running the same race, so to speak. We will refer to this strategy as *marginal inference*. Marginal is a statistical term that means “averaging over;” in assessment the idea is that standardizing all conditions of a test and its items means some particular aspects will favor some students and other aspects will favor others, but these are random differences that average out.

Alternative forms of assessment that include accommodated tests, customized tests, and examinee-choice of tasks propose a different sense of fairness: Tests can differ in their surface characteristics in such ways that equivalent evidence about examinees’ knowledge or skills can be obtained. We will refer to this as *conditional inference*. Conditional is also a statistical term, and refers to drawing inferences where certain information is taken into account specifically as opposed averaged over. In assessment, conditional inference means that aspects of an assessment vary, but are specifically tailored to students so as to enable each individual to access, interact with, and provide responses to tasks in ways that present minimal difficulty, and the primary challenge is the knowledge or skill meant to be assessed.

Assessment that is tailored in one form or another has become widespread, such as the accommodations in testing spurred by the requirements of Americans with Disabilities Act. However, the methodologies of educational assessment and educational measurement (i.e., psychometrics) evolved in the environment of standardized assessment procedures and marginal statistical inference (Green, 1978). 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 present here a framework for assessment design and psychometric modeling that extends familiar

assessment methodology to assessments based on conditional inference. We build on recent work in the three distinct areas that are required jointly to complete the paradigm of conditional inference, namely the theory of assessment design, universal design for learning, and psychometric models. The approach suggests ways of designing accessibility considerations and validity considerations into assessments from the start. Examples are drawn from the Principled Science Assessment Designs for Students with Disabilities project (Haertel et al., 2010), in work supported by the Institute of Educational Sciences, U.S. Department of Education.

1.1 Rationale

Standardized testing first appeared in China more than a thousand years ago. The goal was to provide a basis for evaluating and comparing candidates for civil service across a large and diverse nation. The strategy was to make key aspects of the examination process the same for every examinee, in order to reduce variations in testing procedures that would spuriously advantage some examinees and disadvantage others. Scores on exams under which, unbeknownst to the score user, some examinees had more time than others, for example, or had their work rated by different criteria, are patently unfair. In other words,

Proposition 1: Unidentified nonequivalent surface conditions provide nonequivalent evidence about learners.

Standardized tests tacitly embody a sense of fairness based on making surface conditions for all examinees as close to equivalent as practical. The content of tasks, the format in which they are presented, the requirements for response, and the conditions of performance are all controlled as much as possible so that they are the same for every learner. Accuracy and comparability are achieved at a surface level, in the sense of accurately reflecting what students do in a well-defined, common situation with common evaluation procedures (Rose, Murray, & Gravel, 2012). Such tests have an additional operational advantage. Once a set of tasks, administration conditions, and evaluation procedures are defined and agreed upon, it is straightforward to carry out the procedures. Of course test items differ in the details of the knowledge and skill they demand, in ways that favor some students and disfavor others, but the intention is that these differences

will average out and scores will reflect what the tap in common.¹ The intention is that the assemblage of knowledge, skills, or abilities (KSAs) that are meant to be assessed – the construct, in educational measurement terminology – is implicitly defined by the item construction and test administration processes.

Increased efforts to extend educational experiences, including assessment, to a more diverse population of students call attention to the fact that the same situation need not provide the same learning opportunities to all students; or, in assessment, produce the same information about what they know and can do. As a simple example, 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 one with limited vision. Thus,

Proposition 2: Equivalent surface conditions may not provide equivalent evidence about learners.

Rose, Murray, and Gravel (2012) argue that “to measure underlying constructs accurately requires measurement instruments that are adjustable and flexible enough to be precise in the way that other scientific instruments, like microscopes or binoculars, require adjustment to achieve optimal results for different users” (p. 7). That is,

Proposition 3: Surface conditions that differ in principled ways for different learners can provide equivalent evidence.

In the example of a student with limited vision, the solution is obvious: Provide large print, Braille, a reader, or synthesized speech to convey the word problem in a way that suits the student's capabilities. Then what makes the problem challenging is the arithmetic reasoning, not the knowledge and skill needed to access it. In other situations, the way forward may not be so clear. What kinds of accommodations for limited vision are appropriate in a test of reading comprehension, when decoding text is one of the skills required in the standard version of a text? Is scaffolding appropriate for a multi-step

¹ Generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972) allows for nonequivalent surface conditions by characterizing the effects of differences of surface conditions, and incorporating their effects as uncertainty about marginal inferences to task domains.

science investigation when a student's cognitive capabilities make it difficult for him to manage procedures with many steps? What are appropriate language demands in a history test, when a historical document is difficult to read and some students have limited language skills?

There is no simple or universal answer to these questions. Variants of a task that may be "fair" for one purpose may not be "fair" for another. For example, if the intended use of a reading assessment is drawing inferences from the information in the passage, then all the alternative ways of presenting the text listed above are appropriate. But if the purpose includes decoding print as part of the construct, then providing large print is appropriate but a reader and synthesized speech are not; Braille requires deeper probing of the purpose and the assumptions of the test. *The standard form of the task cannot, in and of itself, provide the information to decide what range of variation will lead to valid inference about the construct*; every task calls upon many KSAs, and which ones are relevant to the construct and which are irrelevant is a matter of intention and purpose.

Once a conditional point of view is assumed, a prospective rather than retrospective stance makes for easier, more defensible, and more explicit assessment design. The key is up-front work in defining the construct and range of ways that access, interaction, and response might be varied so as to make sure that while different students may have different forms of a task, (i) the *construct-irrelevant* demands of the variant each student receives are minimized to the extent possible for each particular student, yet (ii) the *construct-relevant* demands are equivalent for all of them.

1.2 Roadmap of the Paper

Defining and implementing a conditional sense of fairness requires building on developments in assessment design theory, universal design, and psychometric modeling. This article seeks to describe the key ideas of each and how they must come together. Figure 1 illustrates its structure graphically.

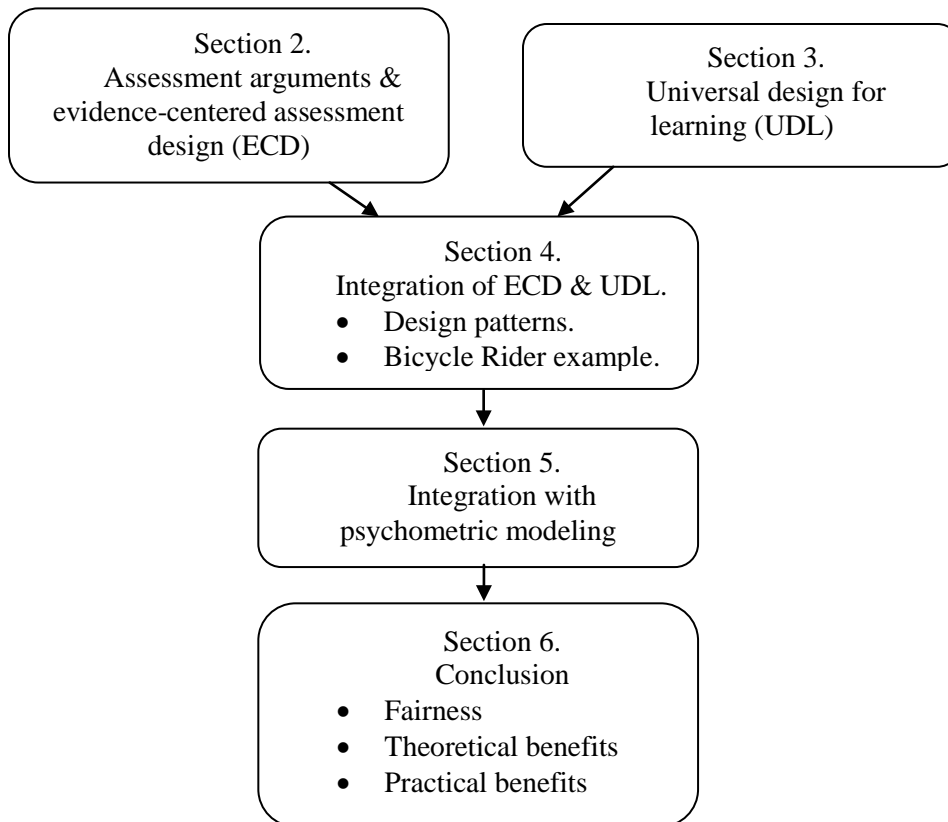


Figure 1: Roadmap of the article

Section 2 reviews the logical assessment-argument structure in which assessment design takes place, as made explicit in the line of research called “evidence centered design” (ECD) (Mislevy, Steinberg, & Almond, 2003; Mislevy & Haertel, 2006). The concept of validity and the role of task variations will be related to this structure. In particular, the interplay among construct-relevant and construct irrelevant KSAs with features of tasks and work products will be explained within the common assessment-argument structure.

Building on research and practical experience with Universal Design for Learning (UDL) (Rose & Meyer, 2002; Rose, Meyer, & Hitchcock, 2005), Section 3 discusses key categories of construct-irrelevant KSAs that hamper students’ learning. Approaches that circumvent, support, or mitigate their detrimental effects in assessment are noted.

Section 4 describes the integration of the ECD and UDL frameworks. This section explicates the essential strategy of accommodation, namely matching students’ capabilities with construct-irrelevant task demands while maintaining construct-relevant

demands (Hansen, Mislevy, Steinberg, Lee, & Forer, 2005; Kopriva, 2008). A support tool for test designers and developers, called a design pattern, which integrates validity principles (Section 2) with UDL principles and techniques (Section 3) is described and illustrated (Haertel, DeBarger, Villalba, Hamel, & Colker, 2010).

Section 5 specifies the psychometric foundations for conditional inference, using the framework of von Davier's (2005) General Diagnostic Model (GDM). The GDM is used to explicate four pertinent evidentiary situations:

- Marginal inference when the testing population is homogeneous with respect to having all the necessary construct-irrelevant KSAs the tasks require. This is the claim made for the traditional standardized testing situation.
- Marginal inference when needed accommodations have not been used and the resulting mismatches are unknown to the score user. This is the situation that accommodations are meant to avoid. For example, score users expect that students with Individualized Educational Plans have received required accommodations during testing situations.
- Conditional inference when task features and student construct-irrelevant capabilities are ascertained after testing occurs. Here students are tested with surface-equivalent forms, but collateral information is available to condition inference after the fact.
- Conditional inference when tasks are matched to students a priori. This is the desired situation when students vary meaningfully with respect to the construct-irrelevant KSAs that are necessary to access, interact with, or respond to assessment tasks.

Section 6 discusses advantages of the approach, both theoretical and practical.

2. Assessment Arguments

ECD is a framework that makes explicit, and provides tools for, building assessment arguments (Mislevy & Riconscente, 2005; Mislevy, Steinberg, & Almond, 2003). Two complementary ideas organize the effort. The first is an overarching conception of assessment as an argument from imperfect evidence. It aims to make explicit the claims (the inferences that one intends to make based on scores) and the nature of the evidence

that supports those claims. The second idea is distinguishing layers at which activities and structures appear in the assessment enterprise. A number of representational forms and tools have been developed to support the work at various layers.

This section briefly describes the ECD layers, enough to coordinate the ideas that are central to the paper: The roles of construct-relevant and irrelevant KSAs in validity, the relationships of these roles to design choices about task features, UDL-infused design patterns as a support tool for task designers, the connection to psychometric models, and the look forward to large-scale implementation of the approach are further detailed. References to fuller discussions and applications are provided for the interested reader.

2.1 Layers in Evidence-Centered Design

Assessment design is often identified mainly with creating tasks. It is advantageous to view the process as first crafting an assessment argument, then embodying it in the machinery of tasks, rubrics, scores, and the like. Messick (1994) succinctly summarizes the core of an assessment argument:

A construct-centered approach would begin by asking what complex of knowledge, skills, or other attribute should be assessed, presumably because they are tied to explicit or implicit objectives of instruction or are otherwise valued by society. Next, what behaviors or performances should reveal those constructs, and what tasks or situations should elicit those behaviors? Thus, the nature of the construct guides the selection or construction of relevant tasks as well as the rational development of construct-based scoring criteria and rubrics. (p. 17)

Evidence-centered design distinguishes layers at which a wide variety of activities and structures appear in the assessment enterprise, all to the end of instantiating an assessment argument in operational processes (Mislevy, Steinberg, & Almond, 2002; Mislevy & Riconscente, 2006). The layers shown in Table 1 are Domain Analysis, Domain Modeling, the Conceptual Assessment Framework (CAF), Assessment Implementation, and Assessment Delivery. They focus in turn on the substantive domain, the assessment argument, the structure of assessment elements such as tasks, rubrics, and psychometric models, the implementation of these elements, and the way they function in an operational assessment.

The key ideas presented in this paper lie in the Domain Modeling and CAF layers. It is in Domain Modeling that assessment arguments are constructed, and we analyze the way that tailoring task features to learners impacts validity. It is in the CAF that the discussion of the corresponding psychometric modeling takes place and particular task features are linked to learners' needs (e.g., perceptual, expressive, cognitive) in an effort to support student performances in non-construct relevant ways. .

Table 1: ECD Layers

ECD 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 Modeling	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	Four-process architecture for assessment delivery systems.

2.2 A Closer Look at the Structure of Assessment Arguments

Messick's quote is a good place to begin understanding assessment arguments, but we need more machinery to examine the effect of task design choices on validity and to provide supports for task developers. We can adapt terminology and representations that Wigmore (1937) and Toulmin (1958) developed for analyzing evidentiary arguments (Mislevy, 2003, 2006).

Toulmin's (1958) schema for how we reason from particular data to claims is shown as Figure 2. A claim is a proposition we wish to support with data. A warrant (possibly multifaceted) justifies the inference from the particular data to the particular claim. In practice we reason inductively, back up through the warrant. We usually have to qualify an inference in light of alternative explanations, which further data might support or undercut.

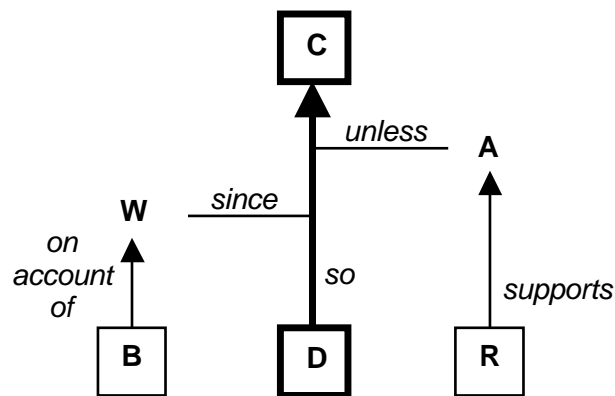


Figure 2: Toulmin's (1958) structure for arguments.

Figure 3 applies the ideas to assessment arguments (Mislevy, 2006). We'll focus on a single task, where a task could be an open-ended problem in a computerized simulation, a language-proficiency interview, a familiar multiple-choice item, or essay question.

The assessment claim is at the top. It is what we would like to say about some aspect of what a learner knows or can do, or doesn't know or can't do, or partially knows at some level or in some way, and so on. At the bottom of the diagram is a student's action in a situation: The student says, does, or makes something. The action in and of itself is not the data, but rather 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.

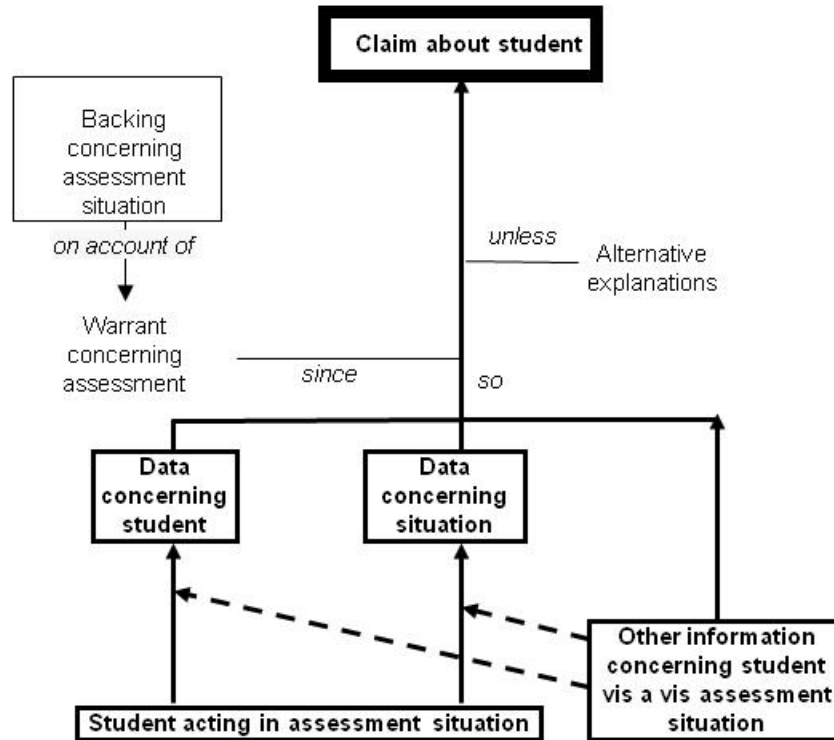


Figure 3: Extended Toulmin diagram for an assessment design argument.

The first of these is usually thought of as “the data” in assessment, and they are indeed the evidence we get from a learner. But features of the situation, the second type of data, are just as critical, in two ways. The task has to have features that engage the KSAs in which we are interested. If the task features present irrelevant impediments to a student’s performance, we don’t get meaningful evidence. 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 barriers.

A warrant about the targeted, or construct-relevant, KSAs, comes with assumptions about access, interaction, and response capabilities. A warrant in an assessment of genetics might look like this:

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 be 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 the crossing and a computer interface to drag and drop the 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 colors that associated with different genetic markers, or cannot physically manipulate the drag-and-drop device, or does not read English well enough to know what is expected. All of these are alternative explanations for poor performance on the standard form of the task, other than the usual claim that she does not understand the genetics. The knowledge, skills and abilities represented in these alternative explanations are construct irrelevant. They represent skills that are required for successful performance, but are not the focus of what the genetics assessment is designed to measure. Presenting the identical task to all students can derail some students as they encounter these construct-irrelevant KSAs and are unable to successfully perform the task. Thus, they may perform poorly on the task, but not because they lack knowledge and skills in genetics. This is what we want to avoid.

2.3 A Closer Look at 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 desired KSAs in a student to the extent the student has them. For example, a simulation task that is meant to get evidence about a student's understanding of Newton's laws but 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 KSAs we care about. A student who can work with Newton's laws but can't figure out the mechanics of the simulation gives "false negative" misleading information. In both cases, alternative explanations, rather than the usual one associated with the warrant, are at play. Messick (1989) calls these threats to validity "construct under-representation" and "construct irrelevant sources of variance." These threats to validity have critical implications for task design.

2.4 Assessment Design Patterns

A design pattern is a formal representation that addresses both a recurring design problem and the core of the solution to that problem in a particular field of expertise. The

design pattern was first introduced in architecture (Alexander, Ishikawa, & Silverstein, 1977) and has been widely adapted in software engineering (e.g., Gamma, Helm, Johnson, & Vlissides, 1994) because of its advantages of reusability and flexibility. A design pattern can be applied repeatedly to resolve a problem in many situations even though the particulars of the situations never remain the same.

The idea of a design pattern was adopted in the PADI project because this project aimed to provide a practical, theory-based approach to developing high-quality assessments of science inquiry (Mislevy et al., 2003). Designing high-quality assessments of science inquiry has been a difficult task largely because it requires the coordination of expertise in different domains – science content experts, science educators and measurement experts. This challenge has been tackled by introducing a design pattern in the assessment design process. In PADI, a design pattern is used as a schema or structure for conceptualizing the components of assessment arguments and their interrelationships. This design pattern plays an important role in bridging content experts and measurement experts so that they can communicate their knowledge in a consistent and effective manner. It also guides assessment designers to think through the essential elements of assessment in ways that lead to a coherent assessment argument and to present their knowledge in a more systematic and fully developed manner. A design pattern can be the basis of principled, even algorithmic, generation of tasks, so that the ideas of UDL adaption discussed below can be incorporated into systems that generate items in real time (Gierl & Haladyna, 2012).

A design pattern contains attributes or constituent pieces of information that address the necessary elements of an assessment argument (Mislevy, 2003). A total of 19 attributes are specified in a design pattern developed in the PADI project, some of which are essential to the assessment argument and some of which are less central. Table 1 provides a list of the key attributes and definition of each, with less central attributes omitted.

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 an inference (*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*). These three attributes are building blocks that the assessment designers should think through during the entire process of task design in order for the assessment argument to be coherent.

Among the other key attributes listed in Table 2, *Rationale* articulates the underlying warrant that justifies the connection between the targeted inferences and the kinds of task and evidence that support them. *Additional KSAs* are other KSAs that may be required in a task that addresses the focal KSAs. Since Additional KSAs are not what are intended to be assessed, they can be potential threats to test validity. Therefore, they first need to be identified and then minimized or avoided in order not to introduce construct irrelevant variance. Alternatively, if it is known that the examinee group of interest possesses sufficient level of a given Additional KSA, the Additional KSA may be incorporated in the assessment tasks along with the intended KSAs. *Potential work products* are students' responses or performances that hold clues or evidence relevant to the Focal KSAs. *Potential rubrics* are links to the rules and instructions that are used to evaluate student work products. *Variable features* are a primary tool for task developers to adjust the difficulty of tasks to focus their evidentiary value on different aspects of the Focal KSA, or to incorporate or circumvent particular additional KSAs. In addition, each design pattern provides links to standards, other design patterns, task templates and exemplary tasks as appropriate.

3. Universal Design Principles

The dialogue around student assessment now encompasses all students as compared to prior compartmentalization that excluded students with disabilities from accountability metrics. From a policy perspective, the definition of "today's students" now includes students who might have previously been exempt from state-level assessments given their special education designation. Beginning with the No Child Left Behind Act in 2001, U.S. states must include students with disabilities in reports of performance and progress. Developing assessment design frameworks that can produce assessment tasks appropriate

and accessible for a wide range of students requires new tools and approaches, including those that can interface with frameworks underlying instructional and assessment materials (i.e., UDL) that *are* specifically designed to meet the needs of students with disabilities.

Table 2. Attributes of Assessment Design Pattern

Attribute	Definition
Title	A short name for referring to the design pattern
Summary	Overview of the kinds of assessment situations students encounter in this design pattern and what one wants to know about their knowledge, skill, and abilities.
Rationale	Why the topic of the design pattern is important.
Focal KSAs	Primary knowledge/skill/abilities of students that one wants to know about for successful performance on the task.
Additional KSAs	Other KSAs that may be required.
Potential observations	Some possible actions that one could observe students doing that would give evidence about the KSAs.
Potential work products	Different modes or formats in which students might produce the evidence.
Potential rubrics	Scoring rubrics that might be useful.
Characteristic features	Kinds 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.
Educational standards	Links to the most 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 give backing for this design pattern.

3.1 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. The first principle, *multiple means of representation*, addresses the ways in which information is presented. The second principle is *multiple means of action and expression*. This principle focuses on the ways in which students can interact with content and express what they are learning. *Multiple means of engagement* is the third principle, addressing the ways in which students are engaged in learning (Rose & Meyer, 2006; Rose & Meyer, 2002; Rose, Meyer, & Hitchcock, 2005). These principles guide the infusion of UDL into assessment design.

Principle I. Provide Multiple Means of Representation (the “what” of learning).

Students differ in the ways that 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), language or cultural differences, and so forth, may all require different ways of approaching content. Some may grasp information best when presented visually or through auditory means rather than the use of printed text alone. Other students may benefit from multiple representations of the content—a print passage presented with an illustrative photographs or line drawings and the use of 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 that 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), those who have language barriers, approach learning tasks very differently and will demonstrate their mastery very differently. Some may be able to express themselves well in text but 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 markedly in the ways in which they can be engaged or motivated to learn. Some students enjoy spontaneity and novelty, while others do not, preferring strict routine. Some will persist with highly challenging tasks while others will give up quickly.

In reality, 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.

3.2 Categories of UDL

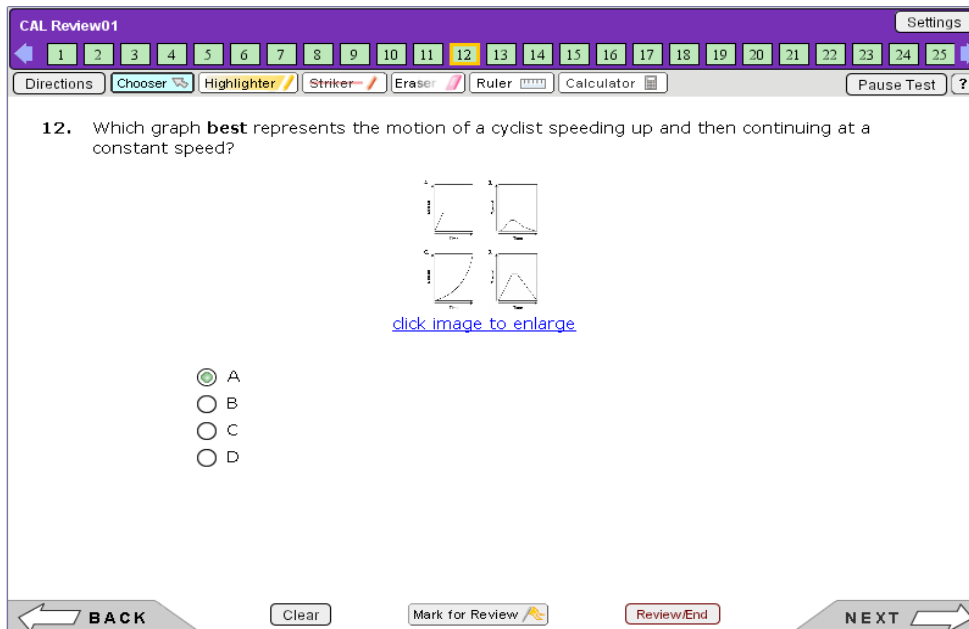
In addition to the three principles of UDL expressed above which provide general guidance for the infusion of UDL into the assessment, we identify particular categories of student abilities (perceptual, expressive, language and symbols, cognitive, executive functioning, and affective) that are required for successful performance on assessment tasks, but are not the assessment targets of interest. We want to use assessment task features that provide supports for students who lack such non-construct relevant abilities, or select features that minimize or eliminate demand for them.

4. Integration of ECD and UDL

In the following section, we present a design pattern that integrates the principles of ECD and UDL. This design pattern illustrates the mechanism by which an assessment designer can infuse explicit principles of UDL into the design of tasks within a domain by identifying non-construct-relevant KSAs and mitigating their influence by designing features of tasks based on UDL principles so as to support or circumvent them. This design pattern is one of a dozen that was created through funding of the Principled Science Assessment Design for Students with Disabilities project (funding from the US Department of Education, Institute of Education Sciences). To help anchor the discussion, we illustrate points with an item called “Bicycle Rider” that was modified with the support of the example design pattern.

4.1 Description of Original Bicycle Rider Item

The original bicycle rider item, shown in Figure 3, is a middle-school assessment item designed to test both an area of science content and an inquiry skill. The science content being assessed is the student’s knowledge of forces and motion in the physical sciences. The inquiry skill concerns the student’s ability to use appropriate tools and technologies to gather, analyze, and interpret data. The item itself describes how a person rides a bike at changing or constant speeds over time. The item then asks the respondent to choose which one of four graphs, each illustrating a different relationship between speed and time, best characterizes the bicycle rider’s travel.

Figure 3. Original Bicycle Rider Item

Original Bicycle Rider Item. Released item in the “7th Grade Science Formative Test” by CAL Testijng (formerly Kansas Computerized Assessments). Retrieved in 2009 from <http://kca.cete.us>, Center for Educational Testing and Evaluation.

The bicycle rider item was taken from a practice test from one state’s large scale middle school science assessment. It is one of 21 discrete, multiple-choice items used in a practice test preparing middle school science students prepare for the statewide science assessment. The original assessment was developed and delivered by the CAL Testing company, and revised, UDL-infused, and field tested in the *Principled Science Assessment Design for Students with Disabilities* project.

The state that developed the original bicycle rider item employed the technology of online assessment for its middle school science assessments. Specific features of the online testing platform included the following (see 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

As developed in its original version for a state-wide assessment, the bicycle rider item is designed to measure two constructs. The first construct is in the physical science content area of forces in motion. The second construct is a science practice that is applicable to all science content areas — namely, the science practice of understanding relationships among data as represented in canonical science and mathematical forms (i.e., tables, charts and graphs). For this single item, then, both the science content and science practice constructs are integrated. Because the item is multiple-choice and scored dichotomously (correct vs. incorrect), the single score can be interpreted to reflect both a student’s abilities in the science content and science practice areas.

4.2 Description of UDL-Infused Design Pattern for the Bicycle Rider Item: Interpreting Data in Tables, Charts, and Graphs

The original version of the bicycle rider item was aligned with a design pattern entitled “Interpreting Data in Tables, Charts, and Graphs.” This design pattern was developed in collaboration with one state department of education for the Principled Science Assessment Design for Students with Disabilities project. Appendix A presents the complete design pattern for Interpreting Data in Tables, Charts, and Graphs. This design pattern supports the writing of items that involve understanding and interpreting data and data variable relationships as represented in tables, charts, or graphic forms. Given that every science content area has the potential to involve data, this design pattern can be used to generate groups of items in all science content areas. Thus, it can be easily used to generate variant assessment items that reduce future design and development costs.

This design pattern also infused principles of universal design for learning (UDL) into specific design pattern attributes. Haertel, DeBarger, Villaba, Hamel, and Colker (2010) provide a more detailed discussion of the integration of UDL into design patterns, 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 that the design

pattern supports. They are intimately connected with the characteristic task features discussed below.

- Additional KSAs are other KSAs that tasks meant to assess the construct may require, and may be either construct relevant or construct irrelevant; it is up to the design to determine which, for the purpose and population of the intended assessment. They are connected with the variable task features, and the design pattern details the relationships.
- Characteristic features of tasks are ones which must be involved in the task somehow if it is to provide evidence about the construct, regardless of other features of the task. Maintaining characteristic features is how the design makes sure that construct relevant KSAs will be probed, no matter how other task features and work products are varied to remove construct-irrelevant KSA demands for them.
- Variable task features include ones that allow a designer to adjust the difficulty, the scope, and the focus of a task while all the while obtaining evidence about the construct. UDL-infused design patterns, in particular, detail features that can be varied to support, circumvent, or appropriately target demands for construct-irrelevant KSAs.
- Potential observations, like characteristic task features, are important to getting evidence about the focal KSAs, regardless of how other features of the task or response may vary. For example, if the necessary evidence in a science task is how a revised model explains observations that were anomalous in an initial model, alternative methods are suggested by which a student can demonstrate the rectified connection between the anomalous data and the revised model (graphical display, verbal explanation, quantitative comparison of residuals from the original and the revised model).
- Potential work products indicate the form in which students can produce responses. They can vary in ways that are sensitive to resource constraints and logistical considerations, and, in UDL-infused design patterns, students' varying profiles for construct-irrelevant KSAs. The Potential work product and Additional KSA categories are linked in such a design pattern to help task

designers see the connections, and in automated task construction systems, to allow for automated task accommodation to learners with different capabilities.

The bicycle rider item was examined in terms of the key design pattern attributes noted above. For example, the primary or focal KSAs to be assessed in the bicycle rider item include:

- 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.

The student behaviors or performances/products that will be accepted as evidence of the KSAs in the bicycle rider item are specified as potential observations and work products. Potential observations include:

- Identification of representational forms of data that communicate the same mathematical relationships among data (or trends in data).

Work products include the

- Selection of an inference or prediction (selected response)

The features of tasks or stimuli that should elicit those cognitive behaviors and performances specified above are presented in the design pattern as characteristic features or variable features. For the bicycle rider item, 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 features intended to influence difficulty of the task are given below. Some of these Variable Features are UDL supports and will be discussed in the following section of this paper. Variable features include:

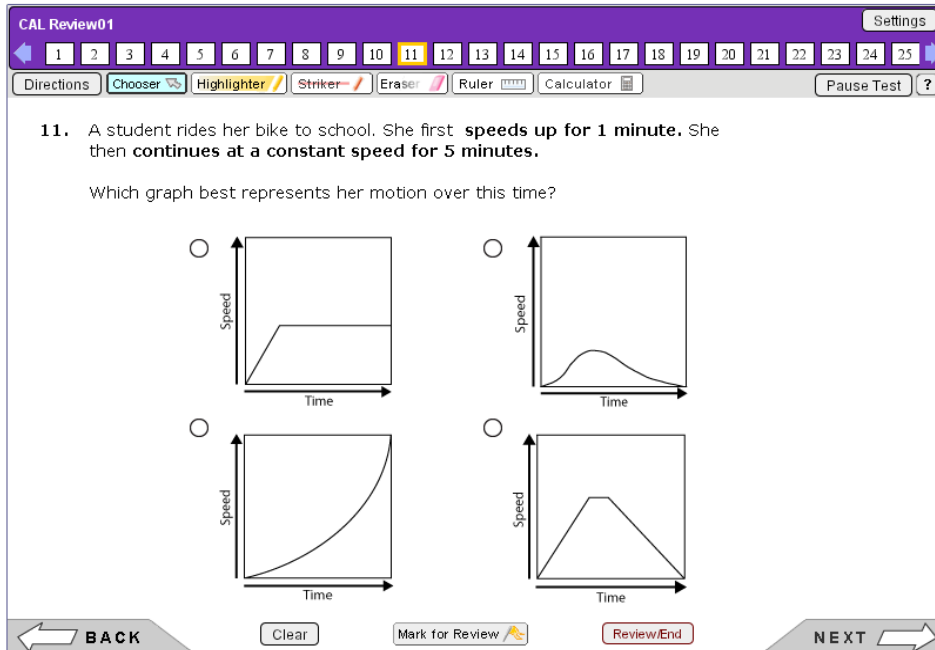
- Number or representations presented
- Types of representations
- Amount of data
- Complexity or representational form(s)

- Number of variables represented in the table, graph, or chart
- Amount of content knowledge required
- Data source (student collected vs. provided)
- Perceptual features: Representational format
- Language and symbols: Supports for vocabulary and symbols
- Cognitive features: Supports for background knowledge
- Cognitive features: Options that guide information processing
- Executive features: Supports for managing information
- Affect features: supports for intrinsic motivation

4.3 Description of Revised Bicycle Rider Item

Figure 4 shows the revision of the original bicycle rider item. By taking the original item and analyzing it in terms of the aligned design pattern, it was possible to identify possible sources of construct irrelevant variance related to individual students' learning needs in terms of perception, expression, language and symbols, cognition, executive functioning, and engagement (affective). These categories of needs were listed as Additional KSAs in the design pattern. Next the Additional KSAs were linked to task model variables (in this case, listed as Variable Features in the design pattern) that could be used to support students' non-construct relevant needs. Then these Variable Features were used to identify a manageable set of UDL-based modifications to the bicycle rider item that potentially could reduce the construct irrelevant variance. These modifications led to a revised version of the original.

Figure 4. Revised Bicycle Rider Item



Modified Bicycle Rider Item. Adapted from a released item in the “7th Grade Science Formative Test” by CAL Testing (formerly Kansas Computerized Assessments). Retrieved in 2009 from <http://kca.cete.us> , Center for Educational Testing and Evaluation.

The specific UDL Principles implemented in the revision of the bicycle rider item are described in Table 3 below. See the Interpreting Data in Tables, Charts, and Graphs design pattern in Appendix A to understand how the UDL features were represented in the design pattern template.

Table 3. UDL Principles (Categories of Students’ Needs) Supported by Variable Features in Bicycle Rider Item

UDL Principle (Category of Student Need)	Task Model Variables Implemented to Address UDL Principles in Bicycle Rider Item
Perceptual Features	<ul style="list-style-type: none"> - Flexible size of text and images, - Flexible amplitude of speech and sound, - Adjustable contrast, - Flexible layout, - Visual graphics, - Verbal descriptors (spoken equivalents for text and images), - Automatic text to speech
Skill and Fluency	<ul style="list-style-type: none"> - Alternative to written response (radio buttons)

Language and Symbols	<ul style="list-style-type: none"> - Embedded support for key terms, - Alternate syntactic levels (simplified text), - Support for decoding (digital text and automatic text to speed)
Cognitive Features	<ul style="list-style-type: none"> - Using explicit examples to emphasize critical concept (minutes cyclist accelerating and at constant speed) - Presentation of graphical representation simultaneously as compared to one at a time (reduce cognitive load)
Executive Features	<ul style="list-style-type: none"> - Reduced working memory - Locate items near relevant text-On-screen - Progress monitoring
Affect Features	<ul style="list-style-type: none"> - Real-world context to heighten engagement - Age-appropriate materials

In comparing the original (Figure 3) and revised (Figure 4) versions of the bicycle rider item, two examples of modifications serve to illustrate the application of UDL via the design pattern attributes. First, note that in the wording of the prompt of the original item, no context is given for the ride or its amount of time. In the revised version of the item, the wording of the prompt references who is riding the bike and the amount of time the ride takes (i.e. adding up to six minutes). This revision was guided by attending to the Cognitive and Affective UDL categories, whereby a real-world context and explicit example of time is added.

Second, note that in the presentation of the original item, the four graph options are presented within an image that needs to be enlarged to be viewed well, and furthermore, the graphs are given a letter (A through D) that must be referenced in order to make the radio button answer choice. In the revision, each of the four graphs is already enlarged (eliminating the enlargement step) and the radio buttons appear directly adjacent to the graphs (eliminating the letter choice translation). This minimizing of extra steps speaks directly to the Skill and Fluency, Cognitive, and Executive Functioning UDL categories. Several of the revision choices were facilitated by the technology platform of that supported the item.

5.0 A Psychometric Framework

This section grounds the conditional-assessment reasoning presented in Section 1.0 of this article from the previous chapters in the terms and models of educational measurement. Section 1 made the philosophical case for a conditional paradigm of fairness. Section 2 laid the assessment design foundations. Section 3 laid the UDL foundations. Section 4 showed how the two frameworks could be integrated, so that tasks that were not identical on the surface could be constructed to evoke comparable evidence even though surface features are varied to tap different levels and combinations of construct irrelevant KSAs, as appropriate to different students so as to minimize irrelevant impediments to their performance. We now lay out a psychometric framework for inference in an assessment designed according to these principles.

In particular, this section shows how the argument structures can be expressed in terms of a psychometric model, namely von Davier's (2005, 2008) General Diagnostic Model (GDM). The GDM is a member of a family of recently-developed class of models called variously cognitive diagnosis models (Leighton & Gierl, 2007; Nichols, Chipman, & Brennan, 1995) and diagnostic classification models (Rupp, Templin, & Henson, 2010). An alternative expression of the ideas in the language of Bayesian inference nets appears in Hansen, Mislevy, Steinberg, Lee, & Forer (2005).

5.1 Key Ideas

A cognitive diagnosis model is a multivariate psychometric model that models probabilities of task response as functions of features of tasks and students' proficiencies with respect to those features. Two features of cognitive diagnosis models that are pertinent to our purposes can be explained by comparing them with more familiar factor analysis models.

Cognitive diagnosis models are similar to confirmatory factor analysis models by allowing the analyst to indicate which proficiencies are involved in a given situation. In factor analysis, the analyst accomplishes this by specifying which variables load on which factors, as suggested by hypotheses about what KSAs are involved in the situations that give rise to the observed variables. In cognitive diagnosis, it appears as indicating

which “attributes” are involved in a given task, while 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, student do or do not have them. This is just right for a domain of binary skills that students must acquire, and are required in various combinations in various tasks. For example, Tatsuoaka’s (1983) analysis of mixed number subtraction characterized students in terms of which mathematical procedures a student has mastered, and tasks in terms of which of those produces they require. The ideas extend readily to more complicated response variables, such as counts, response times, ordered category responses, and continuous measures, and related sets of these, and to more complicated attributes of people, such as ordered and unordered categorical states (e.g., of sophistication of knowledge or level on a learning progression) and continuous variables (such as decoding skill or ability to revise scientific models given familiarity with the model at issue).

Cognitive diagnosis models differ from factor analysis models by allowing a wider range of ways that student proficiencies can be combined to model response probabilities. Whereas factor analysis uses only compensatory combinations – being high in some proficiencies can make up for being low in others – cognitive diagnosis models allow for additional combinations such disjunctions, when different proficiencies can be employed to succeed on a task, and conjunctions, for when certain proficiencies are necessary jointly for high probabilities to succeed no matter how high a student might be on other proficiencies. Conjunctive combinations are exactly what we want for modeling necessary but construct-irrelevant KSAs as in the case of learner needs associated with student disabilities.

Moreover, cognitive diagnosis models allow for logical and probabilistic combinations of these basic structures. We could posit, for example, that in a certain task requires a conjunctive combination of several necessary but construct irrelevant KSAs as well as a set of construct relevant KSAs, which might combine among themselves in various ways but be effective only conditional on sufficiently high values on the conjunction of construct-irrelevant KSAs.

Taking advantage of these flexible combination properties of cognitive diagnosis models, we will describe a basic quantitative model for each of four testing situations that can be described in terms of the qualitative assessment argument structures of the preceding sections. We will use the vector of attribute values to describe which construct-irrelevant and construct relevant demands have been designed into a task variant. We will use the vector of attribute variables to describe students' profiles of construct-irrelevant and construct-relevant KSAs. We will use the structure of the model to express what we think is likely to happen when a particular student is presented a particular variant of a task. We show how this conceptualization fits with the strategy of administering (or custom-building) task variants for each student that give them the best chance to show what they know and can do.

5.2 A General Diagnostic Model

A general form of cognitive-diagnosis type psychometric models is sketched below, and then a particular form is laid out for the purpose of modeling conditional inference as described in this paper.

The key elements are contained in the general form $p(\mathbf{X} = \mathbf{x} | \boldsymbol{\theta}, \mathbf{Q}, \boldsymbol{\eta})$, where $\mathbf{X} = (X_1, \dots, X_n)$ represents n task response variables and $\mathbf{x} = (x_1, \dots, x_n)$ values they can take; $\boldsymbol{\theta} = (\theta_1, \dots, \theta_K)$ is a vector of K KSA proficiency variables that categorize a student (i.e., "attributes" in the cognitive diagnosis literature); \mathbf{Q} is a matrix with n rows, one per task, with the j^{th} row being a vector $\mathbf{q}_j = (q_{j1}, \dots, q_{jK})$ representing features that indicate the qualitative association of Task j to KSAs 1- K ; and $\boldsymbol{\eta}$ a vector of parameters that details the quantitative relationship between between task features and students' probabilities of success. What this expression indicates is that there are multiple aspects of students' knowledge and skill involved; that tasks have features we can relate to these proficiencies in known ways, by virtue of the tasks' construction; and in some manner to be specified, how they interact cognitively determines how students are likely to respond. We can't say more until we assign specific forms to the parameters and the combination functions.

The forms that we will give them in order to make observations about conditional inference are a special case of von Davier's GDM:

- Assume all items are dichotomous, so that $x_j = 1$ indicates a correct response and 0 indicates an incorrect response. (The form of the model generalizes easily to ordered or categorical responses, counts, response times, and multivariate response variables.)
- Partition θ into $(\phi_1, \dots, \phi_k, \theta)$, where the ϕ_k s are construct-irrelevant KSAs and θ is the construct that is the target of measurement. As noted previously, getting evidence from a student about θ inescapably requires some construct-irrelevant KSAs to access, interact with, and respond to a task, and they may be sensory, cognitive, background-knowledge, or other KSAs.
- 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 ϕ s indicate a student's need for such accommodations or modifications. *All* tasks are constructed to have some level of demand on θ . They can differ as to which ϕ_k s they require and what amounts. In particular, we can define families of tasks that are equivalent as to their construct relevant demands, but *differ as to construct-irrelevant demands* (as in Koprika, 2008).
- Define the combination functions $h_k(q_{jk}, \phi_k)$ to take the value 1 if a student's value of ϕ_k equals or exceeds the level of demand for KSA k that is required in Task j , and 0 if not. In other words, $h_k(q_{jk}, \phi_k) = 1$ means that the student is *above the hurdle* with respect to the demands for KSA ϕ_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 ϕ_k , $h_k(q_{jk}, \phi_k) = 1$. When an element of \mathbf{q}_j is defined in terms of the presence of an accommodation or modification, the interpretation is that $h(q_{jk}, \phi_k) = 1$ if either the student does not need the accommodation or modification or if she needs it and it is present. In this case $h(q_{jk}, \phi_k) = 0$ means the student needs the accommodation or modification but

Task j does not provide it. Again $h_k(q_{jk}, \phi_k) = 1$ is interpreted as the student being *above the hurdle*.

- Define the construct-relevant combination function $f(\theta, \beta_j)$ is 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 θ and characteristics of Task j such as its difficulty with respect specifically to θ .
- Let 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, i.e., ϕ_k s. (Together, the π_j s and ϕ_k s constitute the $\boldsymbol{\eta}$ in the general formulation of the GDM.)

The form of the probability model is then given as

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

The term $\prod_k [h_k(\phi_k, q_{jk})]$ is pivotal. By the way that the q_{jk} s, ϕ_k s, and $h(q_{jk}, \phi_k)$ s are defined, this term is $\prod_k [h_k(\phi_k, q_{jk})] = 0$ if there is at least one k for which Task j 's demand with respect to construct-irrelevant KSA ϕ_k exceeds the student's capabilities. In this case the entire second addend is 0; the probability of getting the item right is just π_j , and the response doesn't depend on θ at all! If, on the other hand, for every ϕ_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 $\prod_k [h_k(\phi_k, q_{jk})] = 1$ and the probability of a correct response depends on θ . This is a mathematical form of saying that valid inference about the targeted construct θ is *conditional* on the necessary but construct-irrelevant KSAs the task demands not being appreciable impediments to the student.

This model can be extended in many ways, including alternative response types and multivariate θ s. Another extension would be more gradual h functions. Instead of all-or-nothing, over-the-hurdle-or-not, we could allow performance to gradually degrade as a student fell increasingly below a task's demand for some ϕ_k s.

Putting these ideas into practice requires specifying the structure of \mathbf{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., Rupp, Templin, & Henson, 2010; Kunina-Habenicht, Rupp, & Wilhelm, 2012). The key finding from such research, however, is clear. It is greatly preferable to starting with strong hypotheses from theory and experience, build tasks and accommodation options around these frameworks, then fine tune specifications than to create tasks and try to come up with ϕ s and \mathbf{Q} s after the fact.

It can be fairly argued that the proposed conditional framework introduces a responsibility to test designers and test users to understand the alternative, not-surface-equivalent, forms of tasks do in fact provide equivalent evidence about students. This is so, especially as some instances will not be straightforward. When there is an element of student choice, for example, students can sometimes make choices that disadvantage them (Wainer & Thissen, 1994). When variants differ in terms of the language they are presented, literal translation does not necessarily 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 have been carried out to examine more closely the interacting demands of construct-relevant and irrelevant task features, with an eye toward removing extraneous sources of difficulty for special populations (e.g., 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.²

² It may be noted that it has always been the case that construct-irrelevant demands have unavoidably been present in standardized assessments; it is only recently that they have been recognized as differential threats to validity. Greater awareness and alternative methods bring greater levels of scrutiny to validity threats in new forms, while we have been comfortable in the presence of equal hazards in familiar assessments simply because we are used to the practices.

5.3 GDM Expression of Four Paradigmatic Inferential Situations

We will use the basic form of the GDM for conditional inference, Equation 1, to examine what happens in four different assessment situations, under the assumption that Equation 1 is the correct model.

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

The traditional standardized testing situation, before the introduction of accommodations or modifications, 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. When this is so, $\prod [h_k(\phi_k, q_{jk})] = 1$ for all students, their performance is direct evidence about θ through $f(\theta, \beta_j)$, and, if the everyone-over-all- ϕ -hurdles is correct, it is not even necessary to include the ϕ s and the \mathbf{q} s in the operational model. All of the systematic variation among students' performances is assumed to be due to variation in their θ s – and in this case, the assumption is correct. Familiar scores, whether through classical test theory or IRT, are valid measures of the construct. Under these conditions, equivalent surface conditions do indeed help provide equivalent evidence about students.

5.3.2 *Marginal inference when, unbeknown to the score user, some students are not above some 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 \mathbf{q} features and their consequent ϕ demands, and some students are not above all the hurdles on all the items. In other words, there are at least some items such that for certain students, $h(q_{jk}, \phi_k) = 0$. If tests are scored in the usual way, under the assumption that the preceding case holds instead, then inferences about students' θ s are obtained through $f(\theta, \beta_j)$ for any such item. The student's performance on that item or items, however, is spuriously low, at π_j , and because $\prod [h_k(\phi_k, q_{jk})] = 0$, the response contains no information about her θ .

Note that even if a student is above all the hurdles, so that $\prod [h_k(\phi_k, q_{jk})] = 1$, a student still might not have good chances at getting the item right because her θ might be low – which is what we would like scores to tell us. In other words, getting an item wrong due to lack of some ϕ and getting it wrong because of a low θ have identical observed data, namely, an incorrect response. The difference is that in the first case, the wrong response is misleading evidence about θ , while in the second case it is apposite evidence. We do not want to be in the position of having these alternative potential explanations incorrectly biasing downward our estimates of a student's proficiency. This is patently unfair to the student, and the more likely this situation is to occur in an assessment system, the more its validity is eroded.

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

In this case, students are tested with surface-equivalent forms, but we use a full model something like Equation 1. It is sometimes possible, using the statistical machinery of cognitive diagnosis, to infer students' patterns of ϕ s from the patterns of their responses. Doing so usually requires careful construction of items and tests so the q s are known and properly balanced. If this is done, it is possible to obtain inferences about students' θ s from their response patterns, in effect carrying out *conditional inference by analysis*. In other words, the assessment situation is the same as the second case described above, but now we are using an appropriate psychometric model.

This approach is usually is not very satisfactory, because the added uncertainty that comes from trying to estimate the ϕ s at the same time as θ s renders the estimates quite unreliable. The points to be made from this case, though, are these: It is possible to carry out conditional inference using an appropriate model—not the standard marginal model—and that doing so is generally not practical.

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

This is the situation when (1) students vary meaningfully with respect to the construct-irrelevant KSAs that are necessary to access, interact with, or respond to assessment tasks, (2) we know how they vary, such as by having their IEPs or knowing

what prerequisite knowledge they have, (3) we know or can construct items such that their demands to construct irrelevant KSAs are available, and (4) we assign to each student, for each item, a variant for which $\prod [h_k(\phi_k, q_{jk})] = 1$. We enjoy two benefits.

First, scores depend on θ s, not ϕ s. The assessment is more valid, because many alternative explanations for poor performance due to lack of some necessary ϕ s have been ruled out. Second, because we have done the required work in matching students with task variants, we can again use the simple test scoring models assuming $f(\theta, \beta_i)$ and carry out *conditional inference by design*. The modeling demands are lower, and the reliability as well as the validity of scores is higher, compared to the previous case.

6.0 Conclusions

Below we begin by considering enhancements to the fairness of an assessment, when UDL and ECD are integrated in the assessment design process. In addition, we consider practical benefits of the integration, including increased student engagement and the linking of instruction to student needs.

6.1 Fairness

Achieving fairness in assessment through the integration of ECD and UDL has been a key goal of our work. The 1999 edition of the *Standards for Educational and Psychological Testing* (APA, AERA, NCME) recognized fairness as a fundamental issue of test validity. Our goal to build “fair” assessments is expressed in thoughtfully applying the discipline of ECD in order to provide *all* students with an opportunity to perform at their best in assessment situations. The infusion of UDL into the assessment design from the very beginning is critical to removing barriers that reduce the accessibility of the assessment items and tasks. The *Standards for Educational and Psychological Testing* specifically address the incorporation of UDL as a means for developing tests that are fair to all examinees.

Much of the practice of ECD is focused on the identification of sources of construct-irrelevant variance that can result in faulty interpretations of scores. Assessment design choices that are not carefully examined can contribute to the development of test items that employ unfamiliar language and syntax, poorly understood social and cultural item

contexts and task stimuli, as well as modes of representations (visual, aural, behavioral) that may be systematically biased against subgroups with limited access to requisite background knowledge and use of sensory modalities. Fairness in the assessment situation requires that task contexts be equally familiar, appropriate, and accessible to all students. Articulation of task models from the beginning of the assessment design process reduces the likelihood that items and tasks will be developed that are biased against particular groups.

More recently, with the advent of technology-enhanced assessment delivery systems, students who are unfamiliar with particular hardware and software are at disadvantage in some computer-based testing situations. In particular, those from diverse socio-economic and cultural groups, diverse language backgrounds, and individuals with disabilities need to be considered when technology-based items and tasks are presented.

How does ECD guard against the design of unfair tests? The practice of ECD makes the assessment designer aware of the many kinds of additional KSAs that can contribute to faulty inferences about students' assessment performances. In our work, we consider three broad types of additional KSAs: (1) cognitive background (sometimes referred to as prerequisite knowledge), (2) student needs (perceptual, expressive, language and symbols, cognitive, executive processing, and affective) and (3) technology-related knowledge and skills. As mentioned earlier in this article, the student's needs are identified based on principles of UDL. These needs, if not addressed in the testing situation, can result in a student's poor performance even though she may possess the knowledge and skills of interest.

In applying the ECD process, we identify the focal KSAs that compose the construct we are assessing. Next, the knowledge and skills required to successfully complete an item, but are not the target of the assessment, are identified and labeled as additional KSAs. Then, we reduce the influence of these additional KSAs on a student's assessment performance by identifying variable features that can be designed into the assessment and used to provide non-construct relevant supports. This process of linking the additional KSAs to variable features that support performance without compromising the measurement of the construct of interest guards against inappropriate interpretations of a student's test score.

During the ECD process, we also identify the potential observations needed to provide evidence of whether a student has acquired the knowledge and skills of interest. In articulating these observations, the assessment designer considers whether all students have an adequate opportunity to acquire the knowledge and skills required to perform the focal KSAs. Thus, the role of “opportunity to learn” is prominently considered during the design and development process. By attending to these two processes—identification and mitigation of construct-irrelevant variance and “opportunity to learn”—we increase the fairness of the assessment for all students.

6.2 Theoretical Benefits of Designing Assessments that Integrate UDL and ECD

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. One obstacle to progress has been the many special areas that are involved in this work; few people are experts in all, and there are gaps and conceptual mismatches across workers coming from different backgrounds. The approach demonstrated in this article has the advantage of placing assessment design and analysis within a unified theoretical framework.

A unified framework is important for several reasons. It makes it possible to bring together in a coherent framework the insights and principles that accrue from different fields. A UDL-infused design pattern like the one shown in Appendix A, for example, not only brings in insights from educational technology, science learning research, and universal design for learning, it does so in a form that makes the connections for test designers—and it does so in a way that helps them build tasks that have valid assessment arguments from the very beginning, as well as adapt to the needs of a range of learners. A wealth of knowledge and experience is thus captured in a form that can be used widely by test developers and researchers, cutting across research domains, in common representations and terminology.

The extension to a psychometric framework further aides practice, since there has been little connection between the psychometric community and the UDL community.

The present work articulates the vision of fairness arising from the UDL and special needs communities with the models of performance and formal statistical inference of the psychometric world. The result is a paradigm of fairness that is coherent across these diverse communities, and as such primed for more rapid further development within and across communities.

In particular, this theoretical framework makes it possible to take advantage the opportunities afforded by computer administered testing in a principled way. As Shaftel, Yang, Glasnapp, and Poggio (2005) show, 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. Given prior information about students such as their IEPs, it becomes feasible to envision assembling specific instances of task models to students that may vary in their surface characteristics but be equivalent in the evidence they evoke about the construct of interest (also see Hansen et al., 2005).

6.3 Practical Benefits of Designing Assessments that Integrate UDL and ECD

Each of the benefits below is conferred as a result of the integration of UDL features into the assessment tasks.

6.3.1 Increased Engagement of Students in the Assessment Task

The range of students being tested in accountability situations has increased. In addition, state-of-the-art of assessment design now includes the use of context-rich, situated tasks often presented in online or computer-based testing environments. State-of-the-art tasks often involve story narratives to increase student engagement and motivation and, theoretically, present students with conceptual links previously unavailable in paper-and-pencil testing to support students' cognitive engagement. 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 (e.g., students can fold proteins to create new chemicals to eradicate diseases (Williams, 2009)).

The same characteristics of technology-enhanced tasks that are desirable in terms of assessing students' extended reasoning may present accessibility barriers to students with disabilities. Students with cognitive disabilities, for example, may be overwhelmed with

extended reasoning tasks by virtue of their cognitive load, memory demands, or executive functioning demands. Research has shown that some combinations of stimuli can overwhelm students' working memory. Chandler and Sweller (1992) documented the split attention effect where students' learning was hampered from 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 features that could be implemented to help students monitor their progress could include a progress bar, intermittent messages to the student about their progress, or interactive navigation to support students' working through an extended task.

6.3.2 Linking Instructional Practices to Student Needs

Within domain modeling layer of the ECD process, designers articulate design elements that reflect the assessment of that domain but also reflect aspects of instruction in that domain. Within a domain, designers specify the KSAs including the canonical knowledge representations used in that domain. These are also, in instructional terms, intended learning goals (Krajcik, McNeill, & Reiser, 2008). Designers identify the work products that students would be expected to produce to demonstrate proficiency in a domain. In addition, designers identify qualities of those work products that provide evidence of student understanding, and thereby define the kinds of activities in which students would engage in an instructional context. Because the ECD process has required the identification of additional KSAs, students learning needs have also been identified as well as the particular supports required to support the assessment tasks in the variable features. The linkage among the Additional KSAs and the Variable Features can be applied to, not only performance on the assessment, but also in day-to-day instruction.

A student's performance on a given learning goal can be linked to an additional KSA and supported by the classroom teacher. Additional KSAs provide teachers with information that a student may not have all of the knowledge they need to successfully acquire the new learning goal. For example, additional KSAs reflecting a need for

technology skills may indicate that the teacher needs to provide additional instruction in the use of software and hardware. 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. Using graphic organizers chunking of related content, and making important content salient are all kinds of instructional supports that mitigate the construct-irrelevant variance introduced by unaddressed student needs.

In sum, variable features, articulated in design patterns, take the form of the very same scaffolds that are the critical features of instruction, used to ensure that instructional content is accessible to students. For example, use of multiple representations in instruction can help make instructional concepts salient (Ainsworth, 2006) and might also be used in an assessment design to ensure that focal or target KSAs are presented in multiple ways and remain the primary focus of a task. Similarly, vocabulary support, demonstrations of skills, and contrasting cases might be used in both instructional and assessment contexts. Taken together, the set of variable features defined in the design pattern represent the wide range of supports available in classrooms as well as in assessment context. Although the focus of the present article has been large-scale assessment, the central ideas are equally applicable to classroom testing. Practice is improved by the thought that goes into identifying potential construct-irrelevant demands, and having a set of techniques for reducing them while retaining a focus on construct relevant demands.

ECD provides a set of tools and vocabulary to model the domain of interest, effectively modeling many aspects of the instruction that would be used in a domain. By combining the ECD and UDL frameworks, assessment designs can be linked to, if not embody, the day-to-day instructional contexts of students and address the range of student needs and supports present in classrooms.

7. 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.
- American Educational Research Association. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.
- Ainsworth, S. (2006). DeFT: A conceptual framework for considering learning with multiple representations. *Learning and Instruction, 16*(3), 183–198.
- Alexander, C. Ishikawa, S., & Silverstein, M. (1977). *A pattern language: Towns, buildings, construction*. New York: Oxford University Press.
- 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.
- Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.
- Dolan, R. P., Rose, D. H., Burling, K., Harms, M., & Way, D. (April, 2007). *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.
- Gamma, E., Helm, R., Johnson, R., & Vlissides, J. (1994). *Design patterns*. Reading, MA: Addison-Wesley.
- Gierl, M.J., & Haladyna, T.M. (2012). *Automatic item generation: Theory and practice*. New York: Routledge.
- Green, B. (1978). In defense of measurement. *American Psychologist, 33*, 664-670.
- Haertel, G., DeBarger, A. H., Villabla, 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., 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: 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. *Sci. Ed.*, 92: 1–32. doi: 10.1002/sce.20240
- 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: Macmillan.
- Messick, S. (1994). The Interplay of Evidence and Consequences in the Validation of Performance Assessments. *Educational Researcher*, Vol. 23, No. 2, pp. 13–23.
- Mislevy, R.J. (2006). Cognitive psychology and educational assessment. In R.L. Brennan (Ed.), *Educational Measurement* (Fourth Edition) (pp. 257-305). Phoenix, AZ: Greenwood.
- Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice*. Vol. 25, No. 4, pp. 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., & Riconscente, M. (2005). *Evidence-centered assessment design: Layers, structures, and terminology* (PADI Technical Report 9). Menlo Park, CA: SRI International.

- Mislevy, R. J., Chudowsky, N., Draney, K., Fried, R., Gaffney, T., Haertel, G., et al. (2003). *Design patterns for assessing science inquiry* (PADI Technical Report 1). Menlo Park, CA: SRI International.
- Mislevy, R. J. (2003). Argument substance and argument structure in educational assessment. *Law, Probability and Risk*, 2(4), 237–258.
- Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). Design and analysis in task-based language assessment. *Language Testing*, 19, 477–496.
- Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, 1, 3–62.
- Nichols, P. D., Chipman, S. F., & Brennan, R. L. (Eds.). (1995). *Cognitively diagnostic assessment*. Hillsdale, NJ: Erlbaum.
- 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.
- Rose, D. H., & Meyer, A. (Eds.). (2006). *A practical reader in universal design for learning*. Cambridge, MA: Harvard Education Publishing Group.
- Rose, D. H., & Meyer, A. (2002). *Teaching every student in the digital age: Universal design for learning*. Alexandria, VA: ASCD.
- 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(4), 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: University Press.

- von Davier, M. (2005). A general diagnostic model applied to language testing data. ETS *Research Report RR-05-16*. Princeton, NJ: Educational Testing Service.
- 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.
- Wigmore, J. H. (1937). *The science of judicial proof as given by logic, psychology, and general experience, and illustrated in judicial trials*. (3rd ed.) Boston: Little, Brown.
- Williams, V. (2008). Educational gaming as an instructional strategy. In C. Bonk et al. (Eds.), *Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2008* (pp. 215–220). Chesapeake, VA: AACE.
- Zhang, T., Mislavy, 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.

Appendix A

Design Pattern on Interpreting Data in Tables, Charts and Graphs

[NV] Interpreting Data in Tables, Charts, and Graphs - AERA 2011 | Design Pattern 2130 [[Permit](#) | [Delete](#) | View: View (vertical)]


Title	[Edit] [NV] Interpreting Data in Tables, Charts, and Graphs - AERA 2011
Overview	[Edit] This Design Pattern describes key components of tasks that might be designed to measure students' ability to understand relationships among data as represented in canonical science and mathematical forms (i.e., tables, charts and graphs). Webb's Depth of Knowledge (DOK) framework is used throughout to scaffold design of items that tap this ability at each level of Webb's framework.
Focal Knowledge, Skills, and Abilities	<ul style="list-style-type: none"> FK1. Ability to identify data points in one or more representational forms. FK2. Ability to compare and /or contrast multiple representations and the data represented therein. FK3. Ability to extrapolate or interpolate data points from given data. FK4. Ability to describe simple mathematical relationships or trends among data. FK5. Ability to draw conclusions or make predictions based on data.
Rationale	[Edit] R1. A key activity of science inquiry is working with data in many forms or representations. Students' ability to analyze relationships among data is integral to their participation in science inquiry and to represent and think critically about relationships among experimental variables, observed phenomena, etc.
Additional Knowledge, Skills, and Abilities	<ul style="list-style-type: none"> AK1. ===== The following Additional KSAs are prerequisite knowledge that can be required for tasks that address the Focal KSA. Whether they are to be supported or not (e.g., glossary, background facts, equation list) is a decision to be made -- either by the assessment design team, at the level of the testing program, or at the level of the individual task if that is appropriate in the testing program. ===== AK2. Awareness of different representational forms AK3. Knowledge of what data are AK4. Ability to identify dependent and independent variables AK5. Knowledge of mathematics AK6. Scientific content knowledge AK7. ===== The following Additional KSAs are generally construct-irrelevant knowledge, skills, or other attributes that may be involved in tasks generated under this design pattern. The task author can consider offering supports, presenting material, or getting work products that reduce or avoid requirements for these Additional KSAs, either through accommodated forms of a task or UDL principles. Many of these Additional KSAs are linked to Variable Task Features or Potential Work Products for suggestions on how to do this. ===== AK8. Perceptual <ul style="list-style-type: none"> . vision . hearing . touch AK9. Language and symbols <ul style="list-style-type: none"> . vocabulary and symbols . syntax and underlying structure . English-language proficiency . decoding text or math notation . decoding charts, graphs, or images AK10. Cognitive <ul style="list-style-type: none"> . background knowledge . concepts and categories . information processing strategies . memory and transfer AK11. Skill and fluency


		<ul style="list-style-type: none"> . dexterity, strength, and mobility . navigation and object manipulation . automaticity (e.g., calculations, writing) . familiarity with media . facility with tools
		<ul style="list-style-type: none"> AK12. Executive (problem solving) <ul style="list-style-type: none"> . goal and expectation setting . goal maintenance and adjustment . planning and sequencing steps in a process . managing information and resources . working memory . monitoring progress AK13. Affective <ul style="list-style-type: none"> . intrinsic, task-specific motivation (challenge and/or threat, interest) . sustaining effort and persistence . coping skills and frustration management
Potential observations	[Edit]	<p>Po1. Correct identification of the location of a data point in chart or graph OR the accurate identification of a value to complete a data table.</p> <p>Po2. Identification of representational forms of data that communicate the same mathematical relationships among data (or trends in data).</p> <p>Po3. Accuracy of conclusions drawn from data that are intended to inform predictions.</p> <p>Po4. Appropriateness of inferences drawn from data tables, charts, and graphs.</p>
Potential work products	[Edit]	<p>Pw1. Selection of inference or prediction (selected response)</p> <p>Pw2. Written interpretation of data from one or more representational forms</p> <p>Pw3. Written prediction based on interpretation of data</p>
Potential rubrics	[Edit]	<p>Pr1. Key for selected response items</p> <p>Pr2. Partial credit rubric for scoring of written responses</p>
Characteristic features	[Edit]	<p>Cf1. The presentation contains numeric data</p> <p>Cf2. The presentation includes at least one representational form</p> <p>Cf3. The presented data are in a scientific context</p>
Variable features	[Edit]	<p>Vf1. Number of representations</p> <p>Vf2. Type(s) of representations</p> <p>Vf3. Amount of data</p> <p>Vf4. Complexity of representational form(s)</p> <p>Vf5. Number of variables represented in the table, graph, or chart</p> <p>Vf6. Provision of an example</p> <p>Vf7. Amount of content knowledge required</p> <p>Vf8. Presence of color(s) in table, graph, or chart</p> <p>Vf9. Data source (student collected vs. provided)</p> <p> Vf10. Perceptual Features (1): Representational Format</p> <ul style="list-style-type: none"> - Flexible size of text and images - Flexible amplitude of speech or sound - Adjustable contrast - Flexible colors - Flexible layout <p> Vf11. Perceptual Features (2): Auditory Information</p> <ul style="list-style-type: none"> - Text equivalents (e.g. captions, automated speech to text) - Visual graphics or outlines - Virtual manipulatives, video animation - Verbal descriptions - Tactile graphics, objects <p> Vf12. Perceptual Features (3): Visual Information</p>

- Spoken equivalents for text and images
- Automatic text to speech
- Tactile graphics
- Braille






- VI13. Language and Symbols (1): Supports for Vocabulary and Symbols
- Pre-taught vocabulary and symbols
 - Embedded support for key terms (e.g. technical glossary, hyperlinks/ footnotes to definitions, illustrations, background knowledge)
 - Embedded support for non-technical terms (e.g. non-technical glossary, hyperlinks/ footnotes to definitions, illustrations, background knowledge)
 - Embedded alternatives for unfamiliar references (e.g. domain specific notation, jargon, figurative language, etc.)
- VI14. Language and Symbols (2): Supports for Syntactic Skills and Underlying Structure
- Alternate syntactic levels (simplified text)
 - Grammar aids
 - Highlighted syntactical elements (e.g. subjects, predicates, noun-verb agreement, adjectives, phrase structure, etc.)
 - Highlight structural relations or make them more explicit
- VI15. Language and Symbols (3): Supports for English Language
- All key information in the dominant language (e.g. English) is also available in prevalent first languages (e.g. Spanish) for second language learners and in ASL for students who are deaf
 - Key vocabulary words have links to both dominant and non-dominant definitions and pronunciations
 - Domain-specific vocabulary (e.g. "matter" in science) is translated for both special and common meanings
 - Electronic translation tools, multi-lingual glossaries
- VI16. Language and Symbols (4): Supports for Decoding and Fluency
- Digital text with automatic text to speech
 - Digital Braille with automatic Braille to speech
- VI17. Cognitive Features (1): Supports for Background knowledge
- Advanced organizers, pre-teaching, relevant analogies and examples
 - Links to prior knowledge (e.g. hyperlinks to multimedia, concrete objects in students' environments)
 - Provision of an example
- VI18. Cognitive Features (2): Supports for Critical features, Big Ideas, and Relationships
- Concept maps, graphic organizers, outlines
 - Highlight features in text, diagrams, graphics, and illustrations
 - Reducing the field of competing information or distractions, masking
 - Using multiple examples and non-examples to emphasize critical concepts
- VI19. Cognitive Features (3): Options that Guide Information Processing
- Explicit prompts for each step in a sequential process
 - Interactive models that guide exploration and inspection
 - Graduated scaffolds that support information processing strategies
 - Multiple entry points and optional pathways through content
 - Chunking information into smaller elements, progressive release of information, sequential highlighting
 - Discrete question(s) or scenario-based text presentation
 - Complexity of the scientific investigation presented in the scenario
 - Cognitive complexity (Webb's Depth of Knowledge Levels)
 - If selected response, distractors based on misconceptions/typical errors vs. non-misconceptions
- VI20. Cognitive Features (4): Supports for Memory and Transfer
- Checklists, organizers, sticky notes, electronic reminders
 - Prompts for using mnemonic strategies and devices
 - Templates, graphic organizers, concept maps to support note-taking
 - Scaffolding that connects new information to prior knowledge
 - Embedding new ideas in familiar ideas and contexts, use of analogy, metaphor, example
- VI21. Skill and Fluency (1): Supports for Manipulations
- Virtual manipulatives, Snap-to constraints
 - Nonstick mats, Larger objects
- VI22. Skill and Fluency (2): Supports for Navigation

- Alternatives for physically interacting with materials: by hand, by voice, by single switch, by keyboard, by joystick, by adapted keyboard
- VI23. Skill and Fluency (3): Alternatives to Writing
 - Voice recognition, Audio taping, Dictation, Video, Illustration
- VI24. Skill and Fluency (4): Supports for Composition
 - Keyboarding and alternative keyboards, Onscreen keyboard,
 - Wider lines, Larger paper, Pencil grips
 - Drawing tools - with shapes, lines, etc.
 - Blank tables, charts, graph paper
 - Spellcheckers, calculators, sentence starters, word prediction, dictation (voice recognition or scribe), symbol-to-text, sentence strips
- VI25. Executive Features (1): Support for Goal and Expectation Setting
 - Prompts and scaffolds to estimate effort, resources, and difficulty
 - Animated agents that model the process and product of goal-setting
 - Guides and checklists for scaffolding goal-setting
- VI26. Executive Features (2): Supports for Goal Maintenance and Adjustment
 - Maintain salience of objectives and goals (e.g. reminders, progress charts)
 - Adjust levels of challenge and support (e.g. adjustable leveling and embedded support, alternative levels of difficulty, alternative points of entry)
- VI27. Executive Features (3): Supports for Planning and Sequencing
 - Embedded prompts to "stop and think" before acting
 - Checklists and project planning templates for setting up prioritization, schedules, and steps
 - Guides for breaking long-term objectives into reachable short-term objectives
- VI28. Executive Features (4): Supports for Managing Information
 - Graphic organizers and templates for organizing information
 - Embedded prompts for categorizing and systematizing
 - Checklists and guides for note-taking
- VI29. Executive Features (5): Supports for Working Memory
 - Note-taking, Mnemonic aids
 - Locate items near relevant text
- VI30. Executive Features (6): Supports for Monitoring Progress
 - Guided questions for self-monitoring
 - Representations of progress (e.g. before and after photos, graphs and charts)
 - Templates that guide self-reflection on quality and completeness
 - Differentiated models of self-assessment strategies
- VI31. Affect Features (1): Supports for Intrinsic Motivation (Challenge and/or Threat)
 - Offer individual choice
 - Enhance relevance, value, authenticity (e.g. contextualize to students' lives, provision of an example)
 - Options to vary level of novelty and risk (e.g. options in peer and adult support, alternatives to competition, alternatives to public display or performance, alternative consequences)
 - Options to vary sensory stimulation (e.g. shortened work periods, frequent breaks, noise buffers, optional headphones, alternative settings, presentation of fewer items at a time)
- VI32. Affect Features (2): Supports for Sustaining Effort and Persistence
 - Maintain salience of goals (e.g. explicit display of goals, periodic reminders, replacement of long-term goals with short-term objectives, prompts for visualization)
 - Adjustable levels of challenge and support
 - Encourage collaboration and support
 - Communicate on-going, mastery-oriented feedback
- VI33. Affect Features (3): Support for Self-regulation
 - Guide motivational goal-setting
 - Scaffold self-regulatory skills and strategies
 - Develop emotional self-assessment and reflection

I am a kind of  [Edit]

These are kinds of me  [Edit]

These are  [Edit]

parts of me	
Educational standards	 [Edit] NV (3) Inquiry Standard N.8.A.1 . Students know how to identify and critically evaluate information in data, tables, and graphs
Templates	 [Edit]
Exemplar tasks	 [Edit]
Online resources	 [Edit]
References	 [Edit]

Tags [\[Add Tag \]](#)

(No tags entered.)



Sponsor

The U.S. Department of Education, Grant No. R324A070035

Prime Grantee

SRI International. *Center for Technology in Learning*

Subgrantees

ETS

CAST

