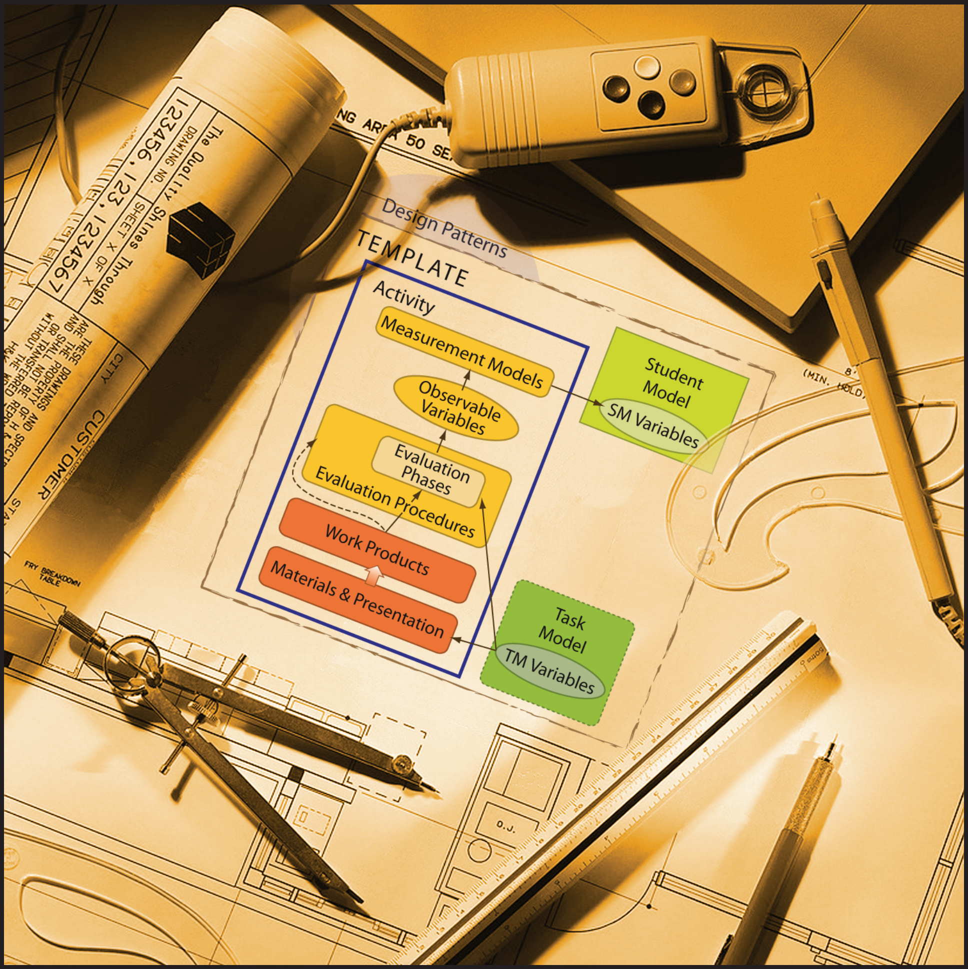


Assessment for Students with Disabilities
Technical Report 1 | June 2010

Using Evidence-Centered Design and Universal Design for Learning to Design Science Assessment Tasks for Students with Disabilities

Project: Principled Science Assessment Designs for Students with Disabilities

Geneva Haertel, Angela Haydel DeBarger, Britte Cheng, Jose Blackorby, Harold Javitz, Liliana Ructtinger, and Eric Snow, SRI International
Robert J. Mislevy and Ting Zhang, University of Maryland
Elizabeth Murray, Jenna Gravel and David Rose, Center for Applied Special Technology
Alexis Mitman Colker, Independent Consultant
Eric G. Hansen, Educational Testing Service

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*

Assessment for Students with Disabilities
Technical Report 1

Using Evidence-Centered Design and Universal Design for Learning to Design Science Assessment Tasks for Students with Disabilities

June 2010

Prepared by:

Geneva Haertel, Angela DeBarger, Britte Cheng, Jose Blackorby,
Harold Javitz, Liliana Ructtinger, & Eric Snow
SRI International

Robert J. Mislevy & Ting Zhang
University of Maryland

Elizabeth Murray, Jenna Gravel, & David Rose
CAST, Inc.

Alexis Mitman Colker
Independent Consultant

Eric G. Hansen
Educational Testing Service

Acknowledgments
This material is based on work supported by the Institute of Educational Sciences, Department of Education under Grant R324A070035 (Principled Assessment Designs for Students with Disabilities).

Disclaimer
Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of IES.

CONTENTS

List of Figures

List of Tables

ABSTRACT

This report describes a design methodology for improving the validity of inferences about the performance of students with disabilities on large-scale science assessments. The work to date combines the use of “universal design for learning” (UDL) with “evidence-centered design” (ECD) to redesign statewide science items to more accurately evaluate the knowledge and skills of all students, including those with high incidence disabilities (mild mental retardation and learning disabilities). The presentation will: (1) describe the state of science assessment for students with disabilities, (2) overview the ECD and UDL frameworks and describe how these frameworks were integrated within a working Web-based assessment design system; (3) describe how the Web-based system helps guide designers through the complex decisions prerequisite to the development of assessments for students with disabilities; and (4) present examples of redesigned science assessment items and design documentation.

1.0 Purpose

The No Child Left Behind Act requires that students with disabilities be included in state assessments and accountability. However, the use of accommodations, modifications, and alternate assessments to permit the inclusion of students with disabilities has given rise to a number of issues related to fairness and test validity. Recently, researchers have begun to explore whether tests can be designed from the outset to be more accessible and valid for a wider range of students; this approach is termed "universal design." The researchers on this project are studying the use of universal design for learning (UDL) paired with an approach termed "evidence-centered design" (ECD) to redesign or develop assessment items that can more accurately evaluate the knowledge and skills of all students on statewide tests. The academic content focus of this study is middle school science, but if successful, the approach can be applied to other topics and age ranges. In this study, the researchers' specific goals are (1) to evaluate the validity of inferences that can be drawn from existing state science assessments for students with and without high incidence disabilities (learning disabilities and mild mental retardation), (2) to redesign assessment items to increase the validity for students both with and without disabilities, (3) to conduct empirical studies of the validity of inferences drawn from the scores on the redesigned items, and (4) to develop research-based guidelines that can be used in large-scale assessment design and development to increase the validity of inferences from science assessment scores for all students.

This paper describes a design methodology for improving the validity of inferences about the performance of students with disabilities on large-scale science assessments. We present work to date from a study that combines the use of “universal design for learning” (UDL) with “evidence-centered design” (ECD) to redesign statewide science items to more accurately evaluate the knowledge and skills of all students, including those with high incidence disabilities (mild mental retardation and learning disabilities). The presentation will: (1) describe the state of science assessment for students with disabilities, (2) overview the ECD and UDL frameworks and describe how these frameworks were integrated within a working Web-based assessment design system; (3) describe how the Web-based system helps guide designers through the complex decisions prerequisite to the development of assessments for students with disabilities; and (4) present examples of redesigned science assessment items and design documentation.

2.0 State of Science Assessment for Students with Disabilities

2.1 Focus on Middle School Science

The decision to focus project research on science assessments for students with disabilities was motivated by the extension of NCLB to science in 2007 and an understanding that success in science coursework serves as a pipeline to scientific careers as well as greater postsecondary education and labor market opportunities for students. Our focus on middle school level students was motivated by the formal introduction of scientific reasoning and problem solving in grades 6 through 8 and by the interdependence of reading, math, and science knowledge, skills, and abilities. Science instruction and assessment are noted for abstract content, challenging vocabulary, text (books) written at difficult readability levels, and complex lab activities. Inability to successfully engage with these curricula and the more complex science content can lead to high school students' decisions to opt out of science classes and scientific career trajectories. Although some special education researchers have developed interventions and outlined best practices for instructing students with disabilities in science (Mastropieri & Scruggs, 1992; Mastropieri & Scruggs, 1995; McClery & Tindal, 1999; Norman, Caseau, & Stefanisch, 1998), science education for students with disabilities has historically been a lower priority in research programs than reading/language arts and mathematics. Moreover, whereas science assessment tasks that entail declarative and procedural knowledge (Li & Shavelson, 2001) require students to recognize and recall information, tasks that entail schematic or strategic knowledge further challenge students to execute or evaluate problem solutions as well as to judge the appropriateness of knowledge applied — precisely the areas affected by many cognition-based disabilities.

2.2 NLTS2 Background

The National Longitudinal Transition Study-2 (NLTS2), funded by the U.S. Department of Education, Institute of Education Sciences (IES), is collecting information from parents, youth, and schools from 2001 to 2010. The study provides a national picture of the educational programs, accommodations, and in–and–out–of–school outcomes of young people with disabilities as they transition from secondary school to early adulthood roles. The NLTS2 sample is comprised of 11,275 students in all disability categories stratified by geographic region and Local Education Agency (LEA) size and LEA wealth. NLTS2 data summaries generalize to the national population of youth with disabilities, as well as to each disability category individually. NLTS2 collects longitudinal data via telephone surveys of parents and youth, paper–based surveys of teachers, and face–to–face assessments of academic performance. To be included in the assessment, students were required to be able to speak and understand English or ASL and be able to complete all measures required for the study using the same accommodations provided to them in the course of day–to–day instruction and assessment. Data presented here are student scores on the version Woodcock Johnson III (WJ3) assessment (Woodcock, McGrew, & Mather, 2001). Data from four subtests were obtained: science concepts, applied problems, passage comprehension, and calculation. Here, we examine the science concepts subtest scores.

Evidence from the WJ3 (Woodcock, McGrew, & Mather, 2001) shows that students with disabilities have difficulties in science in addition to reading and mathematics. On average, students with disabilities nationally score at the 24th percentile on the science concepts subtest, and one in three students has scores below the 5th percentile. Figure 1 illustrates that there is considerable variation in performance both within and across disability categories. This implies that the careful analysis of task requirements, both relevant and irrelevant to the constructs being measured, in this project is appropriate as task requirements may differentially reflect performance of students with different disabilities. For example, in the boxplot below (see Figure 1), students with mental retardation and multiple disability diagnoses have both significantly lower and less variable performance relative to students in other disabilities (for example, learning disability, visual impairment) on the science concepts subtest. This pattern is typical of those in many subject areas tested as part of NLTS2, including the mathematics applied problems subtest that presents students with problems consistent with the types of scientific reasoning commonly introduced in middle school science. It is important to note that the student performance on the WJ3 passage comprehension and calculation subtests are even more variable than those shown below.

Figure 1. Percentile Distribution of Scores on the WJ3 Science Concepts Subtest

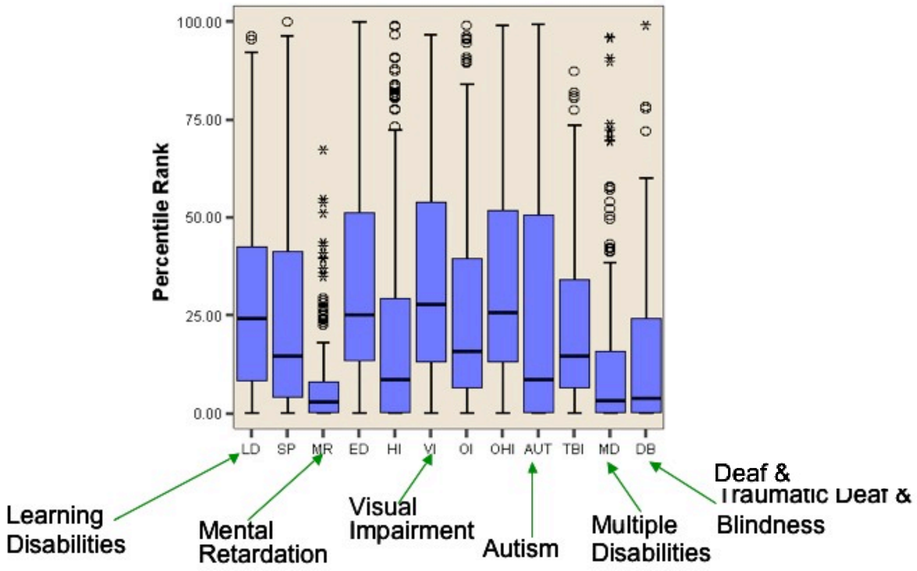


Figure 2. Percentiles on the WJ3 Science Concepts Subtest by Disability Category

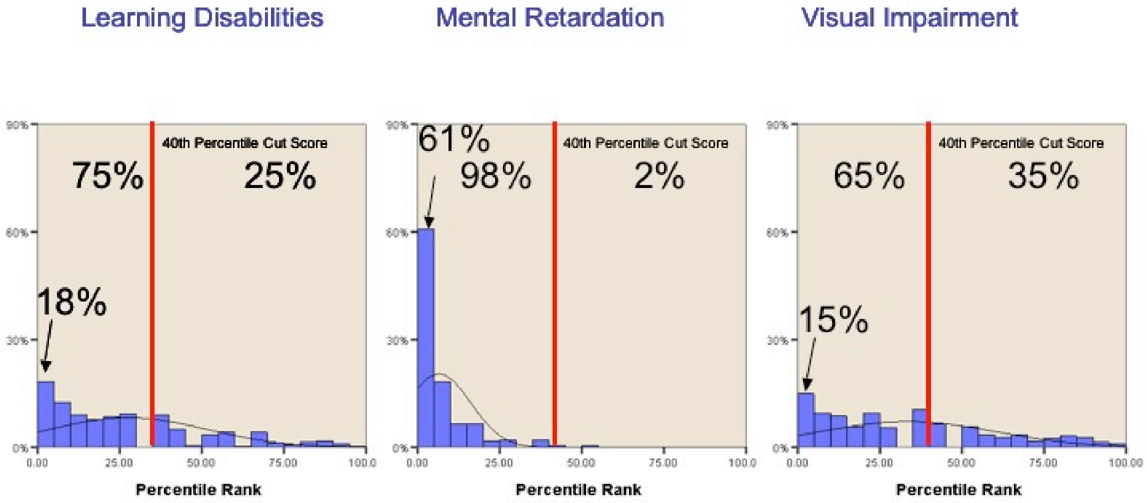


Figure 2 presents an alternative view of these data that illustrates the need to better understand the ways that disabilities differentially limit student performance on assessments. Given that achievement data often are reported in terms of proportions of students above or below a particular threshold, we use this method with a 40 percentile rank threshold to illustrate the variation found in scores on the WJ3 science concepts subtest. This is the way that accountability systems organize achievement data and also provide rough estimates of the level of improvement required for students to meet proficiency targets. This figure illustrates that students with disabilities may not be well represented by traditional reporting of achievement data, due to the highly skewed distribution of achievement scores, although the degree of skewness is highly dependent on students' disability category. Variation across disability categories indicate that to more accurately gauge student performance (and possibly progress over time), we need to understand the variation in student achievement data and, specifically, assessment designers need to document the ways we anticipate task requirements, students' abilities, and their particular disabilities will interact in testing situations. In addition, from a policy perspective it important to explore if some of the achievement gap in test scores can be closed by improvements in test design and administration; for example, by applying universal design for learning (UDL) principles implemented through the PADI online assessment design system — the goal of our project.

3.0 Theoretical Frameworks

The redesign of statewide science items in this project was based on principles of ECD and UDL. In the following section, we describe the principles underlying each of these frameworks.

3.1 Universal Design for Learning

Universal design for learning (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 guiding 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 (Meyer & Rose, 2006; Rose & Meyer, 2002; Rose, Meyer, & Hitchcock, 2005). Described in more detail below, these principles provide structure for the infusion of UDL into assessment.

> **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. Others may simply grasp information better through visual or auditory means rather than printed text. In reality, there is no one means of representation that will be optimal for all students; providing options for representation is essential.
>
> **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, and so forth, approach learning tasks very differently and will demonstrate their mastery very differently. Some may be able to express themselves well in writing text but not oral speech, and vice versa. In reality, there is no one means of expression that will be optimal for all students; providing

options for expression is essential

Principle III: Provide Multiple Means of Engagement (the "why" of learning). There is also an affective 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 engagement that will be optimal for all students; providing multiple options for engagement is essential.

3.2 Evidence-Centered Design

Evidence-centered assessment design (ECD) was formulated by Robert Mislevy, Linda Steinberg, and Russell Almond (2003) at Educational Testing Service. ECD builds on developments in fields such as expert systems (Breese, Goldman, & Wellman, 1994), software design (Gamma, Helm, Johnson, & Vlissides, 1994), and legal argumentation (Tillers & Schum, 1991) to make explicit, and to provide tools for, building assessment arguments that help in both designing new assessments and understanding familiar ones (Mislevy & Riconscente, 2005). Two complementary ideas organize the effort. The first is an overarching conception of assessment as an argument from imperfect evidence. Specifically, it involves making explicit the claims (the inferences that one intends to make based on scores) and the nature of the evidence that supports those claims (Hansen & Mislevy, 2008). The second idea is distinguishing layers at which activities and structures appear in the assessment enterprise, all to the end of instantiating an assessment argument in operational processes. By making the underlying evidentiary argument more explicit, the framework makes operational elements more amenable to examination, sharing, and refinement. Making the argument more explicit also helps designers meet diverse assessment needs caused by changing technological, social, and legal environments (Hansen & Mislevy, 2008). In ECD, assessment is expressed in terms of five layers that provide structure for different kinds of work and information at different stages of the process:

Domain Analysis. In the *domain analysis* layer, research and experience about the domains and skills of interest are gathered—information about the knowledge, skills, and abilities (KSAs) of interest, the ways people acquire KSAs and use them, the situations under which the KSAs are employed, and the indicators of successful application of the KSAs.

Domain Modeling. In the *domain modeling* layer, information from *domain analysis* is organized to form the assessment argument. *Domain modeling* structures the outcomes of *domain analysis* in a form that reflects the narrative structure of an assessment argument, in order to ground the more technical models in the next layer. The PADI online assessment design system uses objects called *design patterns* to assist task designers with *domain modeling*. *Design patterns* play a key role in the present project, as we consider the impact of UDL principles and accommodations on task design and evaluation.

Conceptual Assessment Framework (CAF). The CAF layer concerns technical specifications for operational elements including measurement models, scoring methods, test assembly specifications, and requirements and protocols for assessment delivery. An assessment argument laid out in narrative form at the *domain modeling* layer is here expressed in terms of coordinated pieces of machinery: specifications for tasks, measurement models, scoring methods, and delivery requirements within templates. The central models within the CAF are the Student Model, Evidence Model, and Task Model. In addition, the Assembly Model determines how tasks are assembled into tests, the Presentation Model indicates the requirements for interaction with a student (e.g., simulator requirements), and the Delivery Model specifies requirements for the operational setting. Details about task features, measurement-model parameters, stimulus material specifications, and the like are expressed in the CAF model *templates* in terms of knowledge representations and data structures, which guide their implementation and ensure their coordination. These *templates* are essentially blueprints that specify, at a meta–level, the necessary elements for tasks. The present project will include some work at the CAF layer, as we develop example *templates* that demonstrate how tasks can be developed in accordance with UDL principles and modified in accordance with student needs.

Assessment Implementation. The work in this layer includes activities in preparation for testing examinees such as authoring tasks, calibrating items, finalizing rubrics, producing materials, producing presentation environments, and training interviewers and scorers, all in accordance with the assessment arguments and test specifications created in previous layers of ECD. The ECD approach links the rationales for each layer back to the assessment argument and provides structures that support opportunities for reuse and interoperability.

Assessment Delivery. The work in this layer includes activities such as presenting tasks to examinees, evaluating performances to assign scores, and reporting the results to provide feedback to students themselves, teachers, decision-makers, or other stakeholders.

The ECD framework described in this report applies principles of evidentiary reasoning to handle the complexities of the validity argument (Cronbach & Meehl, 1955; Messick, 1989, 1994; Kane, 1992) associated with accessibility features. The key idea is to lay out the evidentiary structures — or what may be termed the validity argument (or "validation argument" [National Research Council, 2004, p. 104]). An assessment argument can be summarized as comprising: (a) a claim about a person possessing at a given level a certain targeted proficiency, (b) the data (e.g., scores) that would likely result if the person possessed, at a certain level, the targeted proficiency, (c) the warrant (or rationale, based on theory and experience) that tells why the person's level of the targeted proficiency would yield the expected score, and (d) "alternative explanations" for the person's high or low scores (i.e., explanations other than the person's level of the targeted proficiency). The existence of alternative explanations that are both significant and credible might indicate that validity is threatened or being compromised (Messick, 1989).

Much of the analysis that is the focus of this project has to do with these alternative explanations — factors that can hinder an assessment from yielding valid inferences. When the potential for such alternative explanations is recognized at the earliest stages of test design, then later rework and retrofitting can be avoided. The ECD accessibility effort has focused on building argument structures that might help anticipate and address key details of these alternative explanations, particularly as they relate to test takers with disabilities (Hansen & Mislevy, 2008).

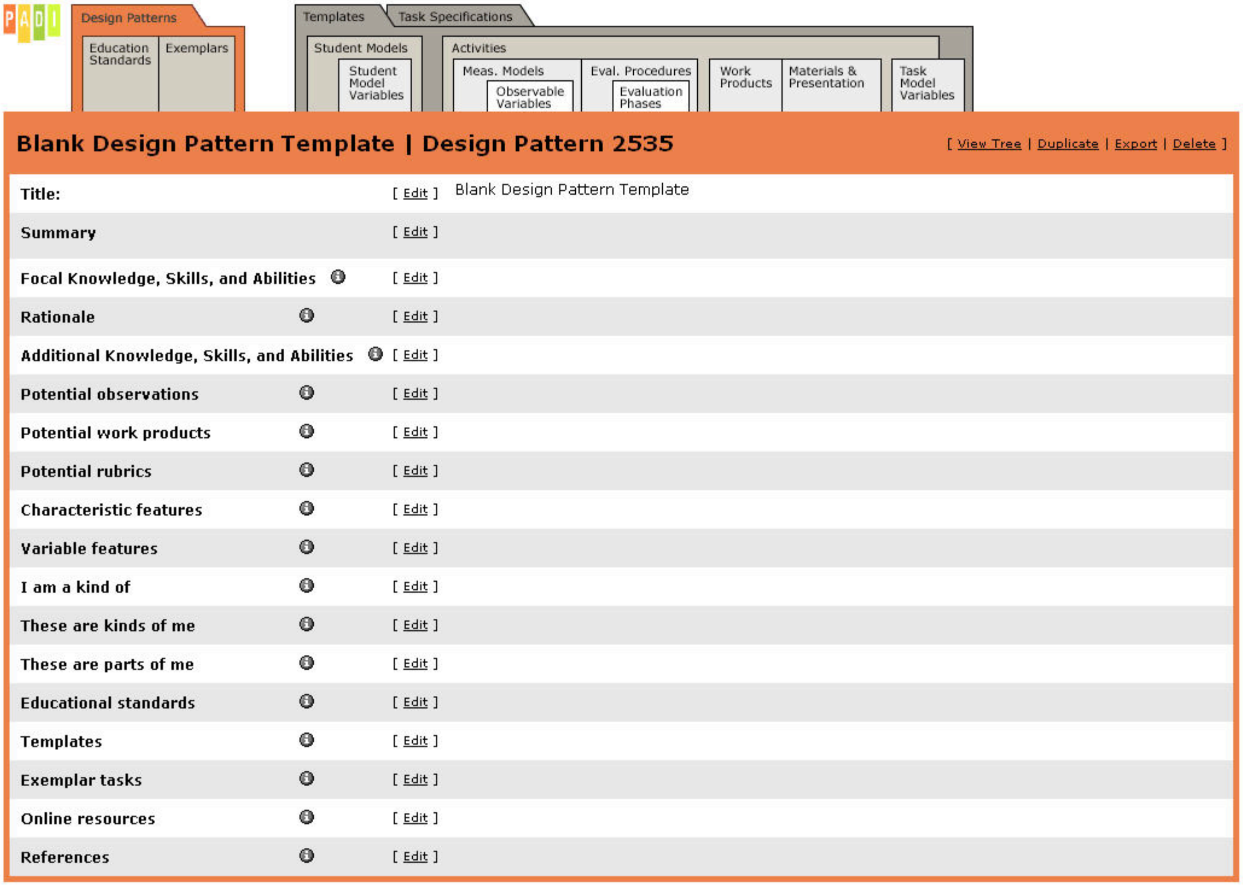
4.0 Integration of UDL and ECD in PADI Online Assessment Design System

Principled Assessment Designs for Inquiry (PADI) was a project supported by the National Science Foundation to improve the assessment of science inquiry (through the Interagency Educational Research Initiative under grant REC-0129331). The PADI project has developed a design framework for assessment tasks based on the evidence-centered design (ECD) framework. PADI was developed as a system for designing blueprints for assessment tasks, with a particular eye toward science inquiry tasks — tasks that stress scientific concepts, problem solving, building models, using models, and cycles of investigation. The PADI framework guides an assessment developer's work through design structures that embody assessment arguments and take advantage of the commonalities across the assessments for sharing and reusing conceptual and operational elements (Mislevy & Haertel, 2006). PADI provides a conceptual framework, data structures, and software supporting tools for this work. The PADI online assessment design system is fully operational.

ECD seeks to integrate the processes of assessment design, authoring, delivery, scoring, and reporting. Work within PADI, however, is focused on design layers that lie above the level of specific environments for task authoring and assessment delivery. The key PADI design objects that will be involved in the present project are *design patterns* and *templates*.

PADI assessment *design patterns* (analogous to those in architecture and software engineering) capture design rationale in a reusable and generative form in the *domain modeling* layer of assessment. They help designers think through substantive aspects of an assessment argument in a structure that spans specific domains, forms, grades, and purposes (Mislevy et al., 2003). Assessment designers working with the PADI design system use the web-based design interface illustrated for *design patterns* (see Figure 3 for the *design pattern* template). Key attributes of a *design pattern* are summarized as follows:

Figure 3. Design Pattern Template



Blank Design Pattern Template | Design Pattern 2535

Title:	[Edit]	Blank Design Pattern Template
Summary	[Edit]	
Focal Knowledge, Skills, and Abilities	[Edit]	
Rationale	[Edit]	
Additional Knowledge, Skills, and Abilities	[Edit]	
Potential observations	[Edit]	
Potential work products	[Edit]	
Potential rubrics	[Edit]	
Characteristic features	[Edit]	
Variable features	[Edit]	
I am a kind of	[Edit]	
These are kinds of me	[Edit]	
These are parts of me	[Edit]	
Educational standards	[Edit]	
Templates	[Edit]	
Exemplar tasks	[Edit]	
Online resources	[Edit]	
References	[Edit]	

Focal KSAs. These are the primary knowledge/skills/abilities of students that one wants to know about and that are addressed by the assessment. Comparability of scores between individuals with and without disabilities is important, which suggests that one should seek evidence about the same set of Focal KSAs, regardless of whether or not the test taker has a disability.

Additional KSAs. These are the other knowledge/skill/abilities that may be required in a task (Mislevy et al., 2003). For tests of academic subjects, the abilities to "see" and "hear" are typically Additional KSAs. On the other hand, for assessment of sight and hearing, respectively, sight and hearing are likely to be defined as Focal KSAs. Notice that there are many disabilities that involve impairments of sight, hearing, or both (e.g., blind, low vision, color-blind, deaf, hard to hear, deaf-blind). Cognitive issues such as dyslexia, attention deficit, and executive processing limitations also can be addressed. Deficits in such Additional KSAs can cause unduly low scores among test takers with disabilities.

Potential Observations. These are possible things that students could say, do, or make that give evidence about the Focal KSAs.

Potential Work Products. These are various modes or formats in which students might produce the evidence relevant to the Focal KSAs.

Characteristic Features. Characteristic Features of the assessment are the features that must be present in a situation in order to evoke the desired evidence about the Focal KSAs (Mislevy et al., 2003).

Variable Features. Variable Features are described as features that can be varied to shift the difficulty or focus of tasks (Mislevy et al., 2003). Variable Features have a particularly significant role with respect to test takers with disabilities and other sub-populations (e.g., speakers of minority language). Much of our attention will be on manipulating Variable Features to reduce or eliminate demands for Additional KSAs in which there may be a deficit while making sure (to the extent possible) that demands for Focal KSAs have not been changed.

5.0 Method

In this study, ECD and UDL were applied to a subset of 20 preexisting statewide science practice items that were developed for online delivery in the state of Kansas. As preparation for the redesign process, *design patterns* were completed that represent the assessment arguments underlying many of the items. Specifically, the *design pattern* helps identify whether task requirements elicit proficiency on intended test constructs (Focal KSAs) or inadvertently contribute variance to student scores but are not relevant to the construct being measured (construct–irrelevant Additional KSAs). Based on this analysis, revisions were made to item designs to reduce the influence of construct–irrelevant Additional KSAs. In the following section, we describe how we implemented revisions to the Kansas practice assessment items.

Construct validity is the sine qua non of assessment properties; to what degree do the evidence and rationale for the data gathered in an assessment support the inferences or decisions that a user wants to make? In the literature on accommodated assessment, the question typically centers on whether a given alteration of a task "changes the construct" (*Standards for Educational and Psychological Testing*, AERA, APA, NCME, 1985. p. 78). Specifically, if an alteration changes the construct, then construct validity has been violated. Conversely, if the alteration does not change the construct, then construct validity has not been violated.

Yet for assessment designers and developers, as well as some other audiences, there is often a need to reason more deeply about the relationships between construct validity and task design. We would argue that it is important to specify more carefully what knowledge and skills need to be assessed and at what levels; the assessment designers need to determine the essence of the intended construct that is to be assessed and what knowledge and skills influence test performance but are not the intended construct. This cannot be determined simply by examining the tasks on a test, because all of the knowledge, skills, and abilities needed to do well on a test are jointly required. In a given testing application, some of these KSAs will be relevant for the inference at hand and others will not (Phillips, 1994); the target examinee population may vary on some of them and not on others. It can even be the case that a given alteration on a test will introduce extraneous score variation in one application, and thus reduce validity, but reduce extraneous variation in a different application of the same test, and increase validity there. It is only by knowing the purpose of a test and the intended examinee population that one can answer how a given change will impact the evidentiary value of data for the construct meant to be assessed. A series of decisions needs to be made in the course of developing a specific test for a specific purpose and testing population to reason through the question of whether a given alteration "changes the construct."

5.1 Infusing UDL into PADI Design Patterns

The project team reviewed relevant background information on ECD and UDL to determine the intersection between UDL principles and PADI *design patterns*. Based on this analysis, six of the original nine UDL categories derived from UDL Principles I, II, and III are now used to categorize types of construct-irrelevant Additional KSAs that are likely to influence student performance. Definitions of UDL categories are provided in Table 1.

The three original Guidelines for Principle I — Perceptual, Language and Symbols, and Comprehension — have been included in the PADI design system. “Comprehension” was changed to “Cognitive” to reflect the types of supports that would be appropriate for assessments. The three guidelines for Principle II were modified for the PADI design system. Physical Action and Expressive Skills and Fluency were combined into one category — Skill and Fluency. The three guidelines for engagement have been condensed into one category — Affect — for the PADI design system. We used these categories to define potential construct-irrelevant barriers to assessment and Variable Features to consider when developing an assessment. These features can be embedded into assessments when appropriate in order to reduce barriers and gain a more accurate understanding of student learning. When using these categories it is essential to keep in mind the knowledge, skills, and abilities being assessed and the impact of any variable feature on the construct relevance of an item.

5.1.1 Variable Feature Categories Derived from UDL Principles

Appendix A lists the Variable Features associated with each UDL category and provides examples. A summary of the features for each category appears below.

Perceptual. To be accurate for a diverse student population, assessments must present information in ways that are perceptible to all students. Perceptual barriers to assessment can be reduced by (a) providing the same information through different sensory modalities, and (b) providing information in a format that can be adjusted (e.g., text that can be enlarged, sounds that can be amplified). Multiple representations such as these can ensure that information is not only accessible to students with particular sensory and perceptual disabilities but also easier to access for many others. When the same information, for example, is presented in both speech and text, comprehension is enhanced for most students. Examples of Variable Features in this category include alternatives for visual information (e.g., providing text-to-speech) and options for representational format (e.g., enlarged text and graphics).

Language and Symbols. Students vary in their facility with language and symbols. Vocabulary that may clarify a test item for one student may be foreign to another. A graph that illustrates the relationship between two variables may be informative to one student but puzzling to another. An important assessment strategy is to ensure that alternative representations are provided not only for accessibility but to ensure that information is clear and understandable to all students. Examples of Variable Features in this category include supports for vocabulary (e.g., definitions of construct-irrelevant terms) and supports for decoding graphs or charts (e.g., providing an explanation of categories in a chart).

Cognitive. Decades of cognitive science research have shown that the ability to transform information into useable knowledge is an active, not passive process. Constructing useable knowledge depends on active "information processing skills" such as selective attending, integrating new information with prior knowledge, and strategic categorizing. Individuals differ greatly in their skills in information processing and in their access to prior knowledge through which they can assimilate new information. Proper design and presentation of information can provide the cognitive ramps that are necessary to ensure that assessments accurately measure student knowledge. Examples of Variable Features in this category include supports for background knowledge (e.g., links to relevant background information) and supports for information processing (e.g., chunking information into smaller elements).

Skill and Fluency. Print format provides limited means of navigation or physical interaction (e.g., turning pages with fingers, writing by hand in spaces provided). Many interactive pieces of educational software similarly provide only limited means of navigation or interaction (e.g., manipulating a joystick, mouse, or keyboard). Navigating and interacting with these materials will raise barriers for some students. It is important to provide assessment materials that students can use easily. Furthermore, no medium of expression is equally suited for all students or for all kinds of communication. Alternative ways to respond to assessment items should be provided in order to ensure that the mode of response does not interfere with students' ability to demonstrate their true understanding. Additionally, students vary widely in their familiarity and fluency with different media. Media used for assessment, therefore, should include supports to scaffold and guide students who are using a less familiar medium so that they can express themselves competently. Examples of Variable Features in this category include supports for composition (e.g., sentence starters) and alternatives to writing (e.g., audiotaping responses).

Executive. Executive functions are at the highest level of the human capacity to act skillfully. We use these functions to overcome impulsive, short–term reactions to our environment and instead to set long–term goals, plan effective strategies for reaching those goals, monitor our progress, and modify strategies as needed. Executive functions have very limited capacity and are especially vulnerable to disability. In assessment situations, students' weaknesses in executive functions can hinder their ability to accurately demonstrate what they know. For these reasons, scaffolding for executive functions is important to consider in assessments. Examples of Variable Features in this category include supports for maintaining a goal (e.g., sentence starters) and supports for planning (e.g., graphic organizers).

Affect. Students differ significantly in what attracts their attention and engages their interest. Even the same student will react differently over time and in different circumstances. Additionally, many tasks require not just initial engagement but sustained attention and effort. When motivated to do so, many students can regulate their attention and affect in order to sustain the effort and concentration that such learning requires. However, students differ considerably in their ability to self-regulate in this way. Examples of Variable Features in this category include supports for intrinsic motivation (e.g., choice of item context) and supports for sustaining effort (e.g., explicit display of a goal, such as number of items completed).

By incorporating UDL into assessment, all students' needs are taken into account. Providing options and supports will reduce potential construct-irrelevant barriers and lead to more accurate measurement for the range of learners who participate in a given assessment. We believe that the careful infusion of the six UDL categories designed for the PADI system will ultimately bring about a greater understanding of student learning.

The six categories within the Additional KSAs, along with the accompanying UDL Variable Features, guide designers to consider the diverse needs of all students. A similar extension of Potential Work Products that would support a range of ways of responding to tasks is being developed and linked with appropriate UDL-motivated KSAs. By infusing UDL into the PADI design system, assessment designers are able to create flexible *design patterns* that will provide a more accurate measure of student learning.

5.2 Background on Design Patterns, Construct Validity, and Specific Assessment Contexts

A *design pattern* helps by laying out choices to be made as appropriate to specific testing applications. It is the specific test and context to which the property of construct validity applies, thus leading to the determination of which potential sources of variance among examinees' test scores would be construct relevant or construct irrelevant.

We have discussed above how important it is for tasks that are intended to assess a Focal KSA to exhibit in some form the Characteristic Features denoted in the *design pattern*, and that by manipulating Variable Features a test developer can increase, decrease, circumvent, or support particular Additional KSAs. A key point is that exactly which Additional KSAs, at which levels, will be construct-relevant to require in a task in a given context, depends on the test purpose and target population; that is, a test–for–purpose–with–population decision. The creator of the *design pattern* does not know what this decision will be, because it can be validly and appropriately different for different applications.

Table 1 below distinguishes between what can be known at the time of creating a *design pattern* for any number of tests that in some way address the Focal KSAs and what must be determined at the time of specifying the application to a particular test. Note that the terms “Focal KSA” and “Additional KSA” describe KSAs in the *design space* while “construct relevant” and “construct irrelevant” describe KSAs in the *application space*.

Table 1. Design Space and Application Space Descriptors

	Application Space Descriptors (for thinking about KSAs for a particular test and its purpose and the intended examinee population)	
Design Space Descriptors (for thinking about KSAs in the *design pattern* stage)	Construct Relevant	Construct Irrelevant
Focal KSA. A *design pattern* is meant to support designing tasks and assessments that assess the Focal KSAs.	(1) Focal KSAs from the *design pattern*, at the right level and focus for the application.	(2) Focal KSAs, but too hard, too easy, or off focus for the intended application.
Additional KSA. Additional KSAs may be required at the designer’s discretion	(3) The designer deems certain Additional KSAs are appropriately part of the intended construct to assess.	(4) Additional KSAs that are required to apprehend, interact with, or respond to an implemented task, yet are not part of the intended construct to assess.

Both Cell 1 and Cell 2 specify what a particular test application needs to require for KSAs that are listed in the Focal KSA attribute of a *design pattern*. A test needs to have some requirement for Focal KSAs (and perhaps some Additional KSAs as well) in order to be valid. Creating tasks that elicit these KSAs will contribute construct relevant variance in examinees’ scores as long as it is done correctly.

Cell 1 addresses an implemented test application’s requirement for the Focal KSAs listed in the *design pattern*, in a task meant to assess the capabilities that the *design pattern* is meant to support, at a level that suits the application’s intended use and examinee population. This is the essence of construct relevant variance in a test: Having the intended capability makes it more likely an examinee will perform well while lacking it makes it more likely that he or she will not perform as well.

Cell 2 concerns requirements for the Focal KSAs described in the *design pattern*, but, in flawed test construction, the demand for the KSAs is not at the right level. For example, the word list for an in-class “spelling bee” for a second grade class might contain words that are much too hard for the students. The KSA of spelling English words is appropriate, but it has not been implemented appropriately for the intended use.

Cells 3 and 4 concern a particular test application’s requirement for KSAs that are listed in the Additional KSA attribute of a *design pattern*. These demands may or may not contribute to construct relevant variance in that application, depending on the purpose and examinee population. In Cell 3, the designer deems that certain Additional KSAs are appropriately part of the intended construct to assess. For example, it may be decided that working memory capability needed to spell words without writing them along the way is appropriate for an in-class “spelling bee” because it is intended to give the students feedback on how well they would do in the upcoming “spelling bee” competition that does not allow writing while spelling. More generally, prerequisite knowledge is often considered “fair game” in assessing school skills. On a test of standards, at a given grade, including requirements for knowledge from standards at earlier grades is often considered appropriate and is a construct-relevant reason for poor performance. Thus, these requirements are not scaffolded (i.e., the Variable Feature “scaffolding” has been set to none for these Additional KSAs).

Cell 4 concerns Additional KSAs that are required to apprehend, interact with, or respond to an implemented task, but which are not part of the intended construct to assess. For example, the standard ‘spelling bee” requires a spoken response, and the KSA of speaking is almost certainly not of the essence for the capability at issue. It is a potential cause of poor performance. Allowing for typed, written, or pointed-to spelling of words as a task feature is a UDL approach to mitigating this problem. In general, requirements in a task for physical and cognitive KSAs that are not construct relevant can lead to poor performance and mask the KSAs that are the intent of assessment (Focal KSAs, plus Additional KSAs, that are construct relevant in the application at hand). They are thus potentially sources of construct irrelevant variation.

Note that for the example of the Additional KSA of being able to say letters aloud — i.e., to produce a Work Product in the form of a spoken sequence of letters — is not universally construct relevant or construct irrelevant in and of itself, but only in light of the purpose of a given test application. The *design pattern* cannot provide the answer, but it can alert the test developer to the question and offer suggests for UDL and accommodation strategies when the Additional KSA is deemed construct irrelevant for the application and there are examinees in the test population who may not have the Additional KSA at the required levels.

The phrase “potentially construct irrelevant sources of variation” highlights the role of the intended examinee population in determining whether a requirement for a construct-irrelevant Additional KSA contributes to invalid inferences in a given application. Being able to speak letters in a “spelling bee" is a construct irrelevant requirement, but if it is known a priori that everyone in the class is able to spell words aloud, this will not be a source of poor performance for this population. But it might be for a different class that has a student who experiences difficulty responding in this manner. An alternative way of responding in that class, perhaps used only by that student, would be necessary in order to remove a construct irrelevant source of variance in the second class.

5.3 Examples of Task Variants

During the first three years of the Principled Science Assessment Designs for Students with Disabilities project, we developed a great deal of machinery for infusing the principles of ECD and UDL into assessments. This includes developing a conceptual framework for the integration of UDL and ECD, extending the PADI *design pattern* tool to build in CAST’s UDL guidelines, co-designing thirteen enhanced *design patterns* with four states, sharing the results at the 2009 annual meeting of the American Educational Research Association (AERA), and setting up arrangements to gather student data on alternate forms of tasks in Kansas and Nevada. As described earlier, the empirical phase of the work has begun with creating revisions for 20 preexisting statewide science items. The revisions represent multiple variations, with each variation crafted to illustrate UDL principles to reduce or remove requirements for construct–irrelevant Additional KSAs for targeted subpopulations of students.

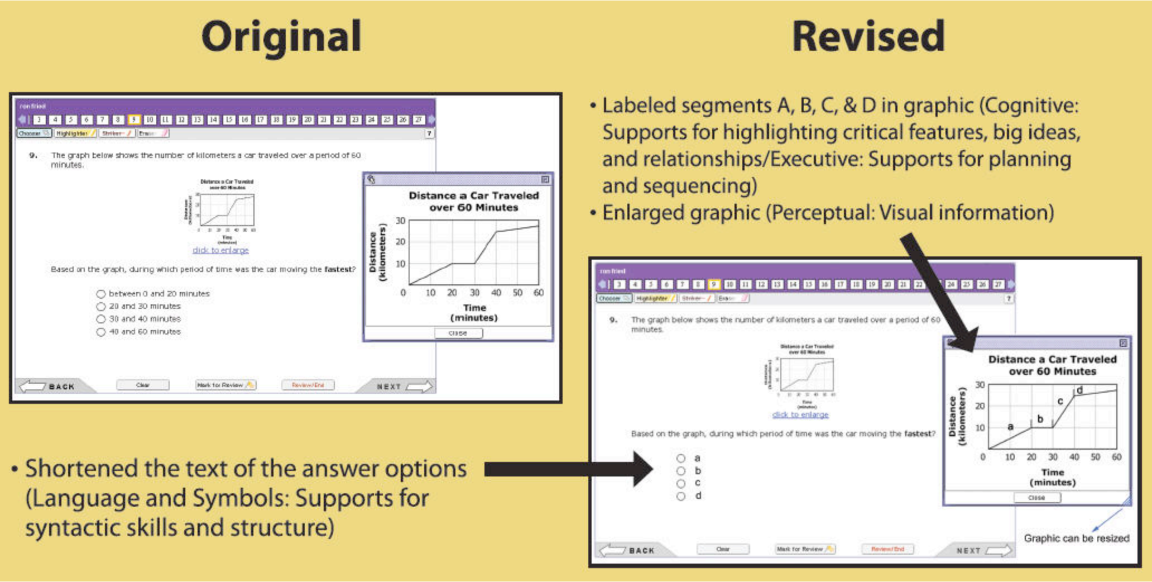
We will be able to administer these variant forms using Kansas’s web-based authoring and delivery system. This delivery system is currently used to administer the Kansas accountability tests to more than 80% of the schools in Kansas. It provides a flexible authoring system that will allow us to implement variant forms of the tasks with little or no additional programming so that we can manipulate presentation material, response modes, supports, and UDL features. This system, in fact, brings a set of UDL features of its own that can be turned on or off to further create task variants.

6.0 Results

We present three examples of original and revised items, including the rationale for the changes made to item directives, graphics, and distracters. The role of technology will be addressed, and features of the online delivery system will be demonstrated.

The tasks that have initially provided the source material for task variants are released tasks from the Kansas assessment and formative assessment tasks, all of which are already implemented in the Kansas statewide delivery system. What follows are discussions of three pairs of tasks that we have developed as initial examples of how we would revise the tasks to take into account the ECD and UDL framework. We provide explications of how the differences in features between the tasks in each pair reflect considerations of UDL and ECD assessment arguments, using the conceptual machinery we have developed in the project. Figures 4 through 6 present three examples of items that have undergone revision using the UDL-infused ECD framework.

Figure 4. Example Item 1: A Time-Distance Graph



Original

9. The graph below shows the number of kilometers a car traveled over a period of 60 minutes.

Based on the graph, during which period of time was the car moving the **fastest**?

○ between 0 and 20 minutes
○ 20 and 30 minutes
○ 30 and 40 minutes
○ 40 and 60 minutes

- Shortened the text of the answer options (Language and Symbols: Supports for syntactic skills and structure)

Revised

- Labeled segments A, B, C, & D in graphic (Cognitive: Supports for highlighting critical features, big ideas, and relationships/Executive: Supports for planning and sequencing)
- Enlarged graphic (Perceptual: Visual information)

9. The graph below shows the number of kilometers a car traveled over a period of 60 minutes.

Based on the graph, during which period of time was the car moving the **fastest**?

○ a
○ b
○ c
○ d

Graphic can be resized

In its original form, the task in Figure 4 presents a graph of the number of kilometers a car traveled over a period of sixty minutes. The graph is a coordinate grid, with the Y-axis representing distance, marked in 10-kilometer ticks, and the X-axis representing time, marked in 10-minute ticks. A piecewise linear line is graphed, with distinguishable segments with different slopes. The prompt asks, based on the graph, “which period of time the car was moving the fastest.” The multiple choice options are “between 0 and 20 minutes,” “20 and 30 minutes,” “30 and 40 minutes,” and “40 and 60 minutes.”

Our analysis of the KSAs demanded by this task included the following:

- seeing the graph (a visual ability)
- "reading" the graph (knowledge and skill requirements to understand the relationship between time and distance, match up the time units in the responses with segments on the X-axis, and relate the verbal descriptions in the options with the line segments on the graph)
- carrying through the steps relating the prompt to the graph (executive processing)
- understanding that the line segment with the steepest slope corresponds to the time period in which the car was moving the fastest (content knowledge).

What KSAs are construct relevant? It is not possible to say by just looking at the task. For the sake of the example, we will presume that the relationship between the slope of a time-distance curve and speed is construct relevant. This being the case, it is necessary that all variants of the task *must* have the properties of communicating to the student that there is a time-distance curve, that there are segments with different slopes, and questioning which segment corresponds to the fastest movement. We have not created a *design pattern* for which this task may be considered an instance. If we had, though, requiring the student to reason through the relationship — the combination of slope and speed — would be a Characteristic Task Feature. Thus, the revised form of the task maintains the feature of reasoning from slopes of the same segments to speed during the time periods.

Is knowledge of the coordinate graph representation construct relevant? In its original form, the task required the student to be able to relate the segments of the line to time intervals, as indicated by the tick marks on the X-axis that are vertically below the endpoints of the segments, and translate this information to the verbal form of the response options. Is this capability construct relevant? Just looking at the item, we do not know. If we had a *design pattern* for interpreting speed-distance graphs, then familiarity with representations would have been listed as an Additional KSA. This means when constructing tasks that elicit information about speed-distance graphs, the task developer must think about how much demand (requirement) for familiarity with representations to impose (the *design pattern* would have given advice for doing this). Should the KSAs at issue be included because it is construct relevant to be assessing them as well as the speed/distance relationship? Should they be avoided because they are construct irrelevant, and we do not know if the students are familiar with them or not? Are they construct irrelevant, but requiring them is appropriate as long as the test provides support for interpreting them? Or are they construct irrelevant but acceptable to require, because it is known that all the students being tested possess these KSAs at the level that will be needed? What kinds of requirements and task demands should be thought through at the task construction phase?

For the sake of this demonstration, we assumed that the capability to identify which time period a line segment represented by its position in the graph was *not* an intended target of inference for this task. Under this assumption, we labeled the line segments a, b, c, and d, in order to eliminate the requirement for identifying line segments by their positions above the X-axis. This would remove a construct-irrelevant source of difficulty and thereby provide *more valid* evidence about the intended KSA for students who lack the ability to identify line segments by their position above the X-axis. A student who answers the revised version incorrectly is providing stronger evidence of the lack of the intended KSA because one alternative explanation of poor performance has been removed.

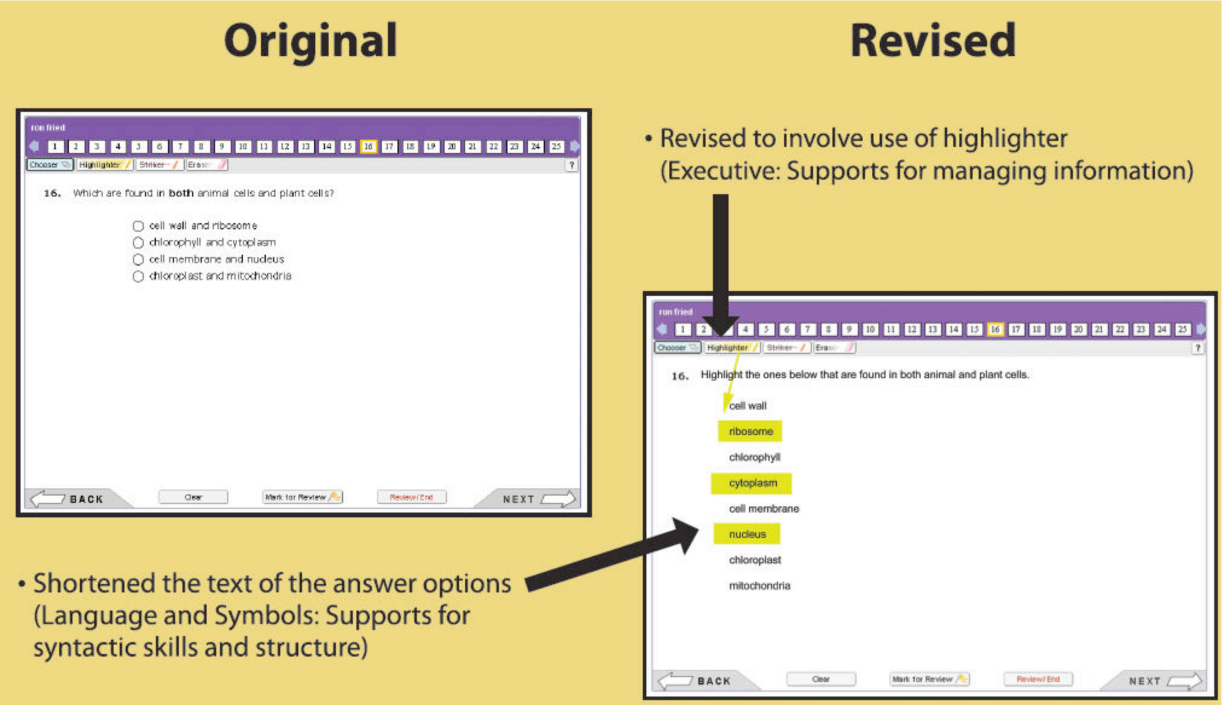
If, on the other hand, the capability to identify which time period a line segment represented by its position in the graph *were* deemed construct relevant, along with the slope–speed relationship, this same modification would have reduced the demand for a construct relevant KSA in the task. The result would be *less valid* evidence for those who lack the ability to identify line segments by their position above the X-axis, because the demand for a construct relevant KSA had been inappropriately removed. Whether this reduction in difficulty increases or decreases validity, however, is indeterminate without determining exactly which KSAs are construct relevant and irrelevant."

In this example, seeing an Additional KSAs in a *design pattern* as including familiarity with representational forms would alert the task developer to determine the extent of demand for this KSA to require in the task, if any. If the determination was that the KSA was construct relevant for the assessment application at hand, then one should maintain this level across revised versions of the tasks that might be administered to different students so that the same mix of construct-relevant KSAs was maintained.

As noted above, the original version of the task requires several steps to reason from line segments to verbal response options, even assuming the student is able to reason through the slope-speed relationship implied by the graph. The revised version of the task simplifies the response options by listing the a, b, c, and d labels of the line segments on the graph. The number and complexity of the steps involved in reasoning from the content issue to response options has been streamlined. For students who understand the key relationship, possible pitfalls in the path of reasoning have been reduced. Does this revision lead to more valid evidence? Although this "after–the–fact" example is not as satisfying as forward design of UDL principles, we would expect that the revision is in fact more valid. In a discussion with state department of education personnel, the question would be this: “Suppose we were having a student talk aloud as he tackled this problem. He explains the relationship between slope and speed and points to the correct time interval, but gets tripped up trying to express this answer in the course of figuring out which response option to check. Would we say that the student really had exhibited evidence of the intended KSAs?” The answer is probably yes. As a rule, revisions prompted by the UDL principles that reduce construct irrelevant sources of cognitive load are usually worth removing if possible.

Since this is a science task, we are probably safe in assuming that “seeing” the graph is not construct relevant, and it is appropriate to reduce or eliminate demands for this ability. One Variable Feature we varied in the revised version of this task, therefore, was providing a larger version of the graph. This revision reduces demands on visual acuity, and makes the task more accessible to students with limited vision. Even the revised version, we may note, has some demand for visual acuity, which would be an alternative explanation for poor performance by a student with little or no visual capabilities. Raised–line graphics and machine–spoken descriptions (via recorded human voice or text-to-speech) would offer possible additional variations of the task that would make it accessible to a still wider range of students.

Figure 5. Example Item 2: Features of Plant and Animal Cells



Original

16. Which are found in **both** animal cells and plant cells?

○ cell wall and ribosome
○ chlorophyll and cytoplasm
○ cell membrane and nucleus
○ chloroplast and mitochondria

- Shortened the text of the answer options (Language and Symbols: Supports for syntactic skills and structure)

Revised

- Revised to involve use of highlighter (Executive: Supports for managing information)

16. Highlight the ones below that are found in both animal and plant cells.

cell wall
ribosome
chlorophyll
cytoplasm
cell membrane
nucleus
chloroplast
mitochondria

In its original form, the task in Figure 5 presents pairs of cell features that may be found in either plant cells, animal cells, or both. The prompt asks students to identify which pair is found in both plants and animals. Note that there are eight features altogether, some of which are found in plant cells, some in animal cells, and some in both.

Our analysis of the KSAs demanded by this task included the following:

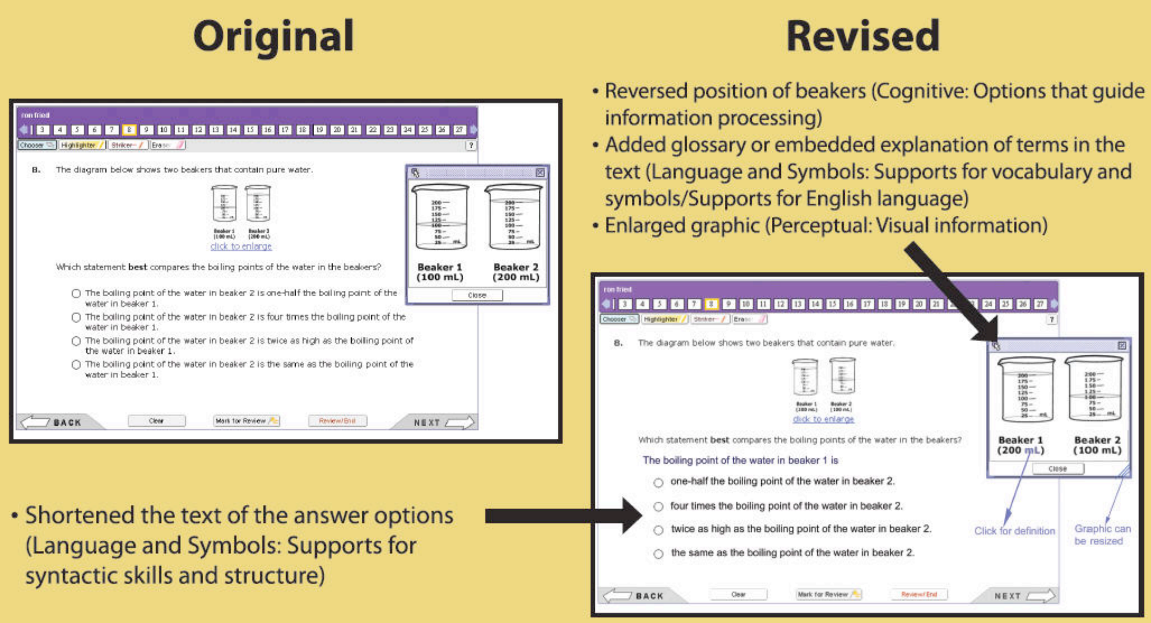
- Seeing the text
- Reading the text
- Knowing which features are in plant cells, which are in animal cells, and which are in both (content knowledge)
- Determining for each pair whether the two items are found in both animal and plant cells (executive function)
- Selecting the correct choice and/or eliminating incorrect ones from among the pairs of features offered (executive function)

Clearly the Focal KSA in this example is the knowledge of which items are found in both plant and animal cells. In the revised form of the task, a student can demonstrate knowledge directly by simply highlighting the correct items. In a pencil and paper format, this item could be completed by circling the correct items. The executive functioning required by the question in its original form is not construct relevant. Note that the revised version untangles the pairs of features and calls for a response concerning each one as opposed to a single response that calls for complex executive functioning that requires the student to both have knowledge of cell features and understand the response format required by the task. The revised version can be scored as partial credit. Not only does the revised version eliminate a construct–irrelevant source of variation due to the requirements of the task format, but it provides more information about the construct–relevant knowledge.

If this task had been created with the support of a *design pattern* based on the Focal KSA, it would be obvious to test developers that the format in the original question imposed construct–irrelevant sources of difficulty that could interfere with accuracy in test results.

If the assessment is to be a multiple choice test, this format may be unavoidable, but it is important to understand the additional construct–irrelevant demands that such a format imposes on students.

Figure 6. Example Item 3: The Boiling Points of Two Beakers of Water



Original

8. The diagram below shows two beakers that contain pure water.

Which statement **best** compares the boiling points of the water in the beakers?

○ The boiling point of the water in beaker 2 is one-half the boiling point of the water in beaker 1.
○ The boiling point of the water in beaker 2 is four times the boiling point of the water in beaker 1.
○ The boiling point of the water in beaker 2 is twice as high as the boiling point of the water in beaker 1.
○ The boiling point of the water in beaker 2 is the same as the boiling point of the water in beaker 1.

- Shortened the text of the answer options (Language and Symbols: Supports for syntactic skills and structure)

Revised

- Reversed position of beakers (Cognitive: Options that guide information processing)
- Added glossary or embedded explanation of terms in the text (Language and Symbols: Supports for vocabulary and symbols/Supports for English language)
- Enlarged graphic (Perceptual: Visual information)

8. The diagram below shows two beakers that contain pure water.

Which statement **best** compares the boiling points of the water in the beakers?

The boiling point of the water in beaker 1 is

○ one-half the boiling point of the water in beaker 2.
○ four times the boiling point of the water in beaker 2.
○ twice as high as the boiling point of the water in beaker 2.
○ the same as the boiling point of the water in beaker 2.

In its original form, the task in Figure 6 presents a picture of two same-sized glass beakers, side by side. The one on the left is labeled “Beaker 1 (100 mL),” and the one on the right is labeled “Beaker 2 (200 mL).” There appears to be a clear liquid in each beaker. Just as suggested by the labels, Beaker 1 is filled to a line marked in on the beaker as “100” and Beaker 2 is filled to a line marked “200.” The term “mL” does not accompany each line of the marked (graduated) beakers, but the term is shown near the bottom of each beaker some distance to the right of the line marked “25” (i.e., the 25 milliliter mark).

Our analysis of the KSAs demanded by this task included the following:

- Seeing the text
- Seeing details of the picture
- Knowing how to read and comprehend complex syntax
- Easily process options where the mention of beakers in the options differs from the order in the picture
- Knowing the meaning of "mL"
- Knowing the meaning of boiling point for a liquid (for example, that it has nothing to do with a point or position on a beaker)

In this example, the Focal KSA is knowing the meaning of boiling point for a liquid (for example, that it has nothing to do with a point or position on a beaker). However, several aspects of this item may result in errors due to variables that are not relevant to this construct. As the image provided with the item presents essential information, providing a way for students to enlarge it so that they can clearly see this information again may decrease the likelihood that some students would score incorrectly due to a factor that is not construct relevant.

The response options each begin with the same eight word phrase. In the revised version this phrase appears as a stem, with only the relevant information in the options. This lowers the reading and language demands of the item while not impacting the measurement of the Focal KSA. Similarly, the response options refer to the beakers in the reverse order of their appearance in the image, a point that some students may miss, resulting in an error not due to poor understanding of the Focal KSA. The order of the beakers has been reversed in the revised item to avoid this.

Another consideration for test items is vocabulary. As illustrated in this example, terms and abbreviations such as “mL” may not be construct relevant, and, if so, their definitions could be given in a glossary. Alternatively, if vocabulary terms are determined to be construct relevant, then of course this support would not be provided.

Thus, in each of the above, the task revision results in reducing a requirement for a construct–irrelevant KSA such that the student’s lack of those abilities should not hinder him or her from demonstrating ability in the targeted proficiency (which, in this case has one KSA). Essentially, revisions to the task ensure that the student can satisfy requirements for all construct–irrelevant KSAs, thereby enabling his or her ability in the construct–relevant KSAs to be expressed.

Of course, for any number of reasons, the student may lack the construct–relevant KSA and, therefore, may perform poorly. However, the assessment will still be valid, i.e., validly low. And with assessments that are more valid for students with disabilities, we will more likely be able to address the range of causes for low academic achievement in many students with disabilities.

7.0 Significance

This application of UDL principles to the revision of statewide assessment tasks systematically documents the integration of UDL and ECD frameworks to enhance construct validity. Our findings speak in particular to those interested in building assessment design arguments that address the issue of student diversity. In the final year of the project, design strategies will be identified that proved especially helpful in the improvement of items for all students and, in particular, for students with high incidence disabilities. These design strategies will be articulated in a set of guidelines that state departments of education can apply in developing items for their statewide science assessments. While these strategies were developed in the context of middle school science assessments, we believe they will be applicable to most subject areas and across students of different ages. Likewise, even though this project focused on the refinement of items in “general education” science assessments, the Variable Feature categories based on UDL principles include features that are appropriate for students with low incidence disabilities.

References

American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education. (1985). *Standards for educational and psychological testing.* Washington, DC, American Psychological Association.

Breese, J. S., Goldman, R. P., & Wellman, M. P. (1994). Introduction to the special section on knowledge-based construction of probabilistic and decision models. *IEEE Transactions on Systems, Man, and Cybernetics, 24*, 1577-1579.

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

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.

Hansen, E. G., & Mislevy, R. J. (2008). *Design patterns for improving accessibility for test takers with disabilities (RR-08-49).* Princeton, New Jersey: ETS Research Report Series.

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

Li, M., & Shavelson, R. J. (2001). (April 12, 2001). *Examining the links between science achievement and assessment.* Paper presented at the annual meeting of the American Educational Research Association, Seattle.

Mastropieri, M. A., & Scruggs, T. E. (1992). Science for students with disabilities. *Review of Educational Research, 62*, 377-411.

Mastropieri, M. A. & Scruggs, T. E. (1995). Teaching science to students with disabilities in general education settings. *Teaching Exceptional Children*, *27*(4). pp. 10-13.

McCleery, J. A., and Tindal, G. A. (1999). Teaching the scientific method to at-risk students and LD students through concept anchoring and explicit instruction. *Remedial & Special Education* 20(1): 7–18.

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*, *32*(2), 13-23.

Meyer, A., & Rose, D. (2006). Preface. In D. Rose, & A. Meyer (Eds.), *A practical reader in universal design for learning.* (pp. vii-xi). Harvard Education Press, Cambridge, MA

Mislevy, R. J., & Haertel, G. (2006). Implications for evidence-centered design for educational assessment. *Educational Measurement: Issues and Practice, 4,* 6-20.

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

Mislevy, R., & Riconscente, M. (2005). *Evidence-centered assessment design: Layers, structures, and terminology (PADI Technical Report 9).* Menlo Park, CA: SRI International.

Mislevy, R., Hamel, L., Fried, R., G., Gaffney, T., Haertel, G., Hafter, A., Murphy, R., Quellmalz, E., Rosenquist, A., Schank, P., Draney, K., Kennedy, C., Long, K., Wilson, M., Chudowsky, N., Morrison, A., Pena, P., Songer, N., Wenk, A. (2003). *Design patterns for assessing science inquiry* (PADI Technical Report 1). Menlo Park, CA: SRI International.

National Research Council (2004) Keeping score for all: The effects of inclusion and accommodation policies on large-scale educational assessment In J. . A. Koenig & L. F. Bachman (Eds). Committee on Participation of English Language Learners and Students with Disabilities in NAEP and Other Large-Scale Assessment. Washington, DC: National Academy of Sciences.

Norman, K. I., Caseau, D., and Stefanich, G. P. (1998). Teaching students with disabilities in inclusive classrooms: Survey results. *Science Education* 82: 127–146.

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

Rose, D. H., & Meyer, A. (2002). *Teaching every student in the Digital Age: Universal Design for Learning*. Alexandria, VA: ASCD.

Rose, D. H., Meyer, A., & Hitchcock, C. (2005). *The universally designed classroom: Accessible curriculum and digital technologies.* Cambridge, MA: Harvard Education Press.

Tillers, P. and Schum, D. A. (1991) A theory of preliminary fact investigation. *University of California Davis Law Review, 24*, pp.931 - 1012.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson III*. Itasca, IL: Riverside Publishing.

Appendix A:
Variable Features by UDL Category with Examples

Perceptual Features

1. Supports for Representational Format
   - Flexible size of text and images
   - Flexible amplitude of speech or sound
   - Adjustable contrast
   - Flexible colors
   - Flexible layout

2. Supports for Auditory Information
   - Text equivalents (e.g. captions, automated speech to text)
   - Visual graphics or outlines
   - Virtual manipulatives, video animation
   - Verbal descriptions
   - Tactile graphics, objects

3. Supports for Visual Information
   - Spoken equivalents for text and images
   - Automatic text to speech
   - Tactile graphics
   - Braille

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.)

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

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

4. Supports for Decoding and Fluency

   - Digital text with automatic text to speech
   - Digital Braille with automatic Braille to speech

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

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

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, distracters based on misconceptions/typical errors vs. non-misconceptions

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

Skill and Fluency

1. Supports for Manipulations
   - Virtual manipulatives, Snap-to constraints
   - Nonskid mats, Larger objects

2. Supports for Navigation
   - Alternatives for physically interacting with materials: by hand, by voice, by single switch, by keyboard, by joystick, by adapted keyboard

3. Alternatives to Writing
   - Voice recognition, Audio taping, Dictation, Video, Illustration
   - (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

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

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)

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

4. Supports for Managing Information
   - Graphic organizers and templates for organizing information
   - Embedded prompts for categorizing and systematizing
   - Checklists and guides for note-taking

5. Supports for Working Memory
    - Note-taking, mnemonic aids
    - Locate items near relevant text

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

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)

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

3. Support for Self-regulation
    - Guide motivational goal-setting
    - Scaffold self-regulatory skills and strategies
    - Develop emotional self-assessment and reflection

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

Prime Grantee
SRI International. *Center for Technology in Learning*

Subgrantees
University of Maryland
CAST