

Assessment for Students with Disabilities
Technical Report 4 | June 2012



UDL and the PADI Process: The Foundation

Project: Principled Science Assessment Designs for Students
with Disabilities

David Rose, Elizabeth Murray, and Jenna Gravel, CAST

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 © 2012 SRI International. All Rights Reserved.

Technical Report 4

UDL and the PADI Process: The Foundation

Prepared by:
David Rose, CAST
Elizabeth Murray, CAST
Jenna Gravel, CAST

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.

Table of Contents

Abstract.....	3
1.0 Introduction.....	3
2.0 Origins of Universal Design and Universal Design for Learning.....	3
2.1 The Basic Framework of UDL	4
3.0 UDL and Large–Scale Assessment.....	5
3.1 UDL and the Optimization of Validity.....	6
3.2 UDL and the Maximization of Accuracy.....	7
3.3 Applying UDL to Large–Scale Assessments.....	7
3.3.1. Category One: Perceptual	8
3.3.2. Category 2: Language and Symbols	10
3.3.3. Category 3: Cognitive	12
3.3.4. Category 4: Skill and Fluency.....	14
3.3.5. Category 5: Executive Functions	16
3.3.6. Category 6: Affective.....	18
4.0 Conclusion	19
References.....	21
Figure 1. Skateboarder Investigation Item Original	9
Figure 2. Skateboarder Investigation Item Revised.....	10
Figure 3. Sugar Dissolved in Water Item Original	11
Figure 4. Sugar Dissolved in Water Item Revised.....	12
Figure 5: Magnets Item Original.....	13
Figure 6: Magnets Item Revised.....	14
Figure 7: Calculation of Force Item Original	16
Figure 8: Toy Truck Experiment Item Original.....	17
Figure 9: Toy Truck Experiment Item Revised	18
Figure 10. Bird Nest Item Original	19

Abstract

This report addresses the embedding of principles of Universal Design for Learning (UDL) into the PADI (Principled Science Assessment Design for Students with Disabilities) process, focusing on large-scale assessment. The key idea in applying UDL principles to assessment is to provide each student with a form of a task for which construct irrelevant, or undesirable, sources of difficulty are minimized for that student, so that his or her performance depends to the largest extent possible on the construct relevant demands of the task. A task may be presented in different surface forms to different students in order to reduce construct-irrelevant demands for each of them, but the construct-relevant demands should be equivalent in each form of an item. This report presents background on UDL and its integration into the PADI process, and examples of its application to assessment items.

1.0 Introduction

During the last few years, the field of Universal Design for Learning (UDL; Rose & Meyer, 2002; Rose, Meyer, & Hitchcock, 2005) has emerged as an important element of educational reform. The recent definition of UDL in congressional legislation (e.g. The Higher Education Opportunity Act of 2008) and inclusion in the Common Core Standards (2010) are but two recent indicants of its increasing visibility. Most recent textbooks on special education, and even regular education, include major treatments of UDL or are entirely structured around the principles of UDL.

In spite of this increasing visibility, however, the implementation of UDL is still in its infancy. Most educators really do not know what “universal design for learning” means, and there are few fully realized applications of its principles in practice. Nowhere is this more evident than in assessment. While there is frequent reference to Universal Design for Learning in the field of assessment, there is considerable lack of clarity as to what the term implies and far too little research available on its effects.

The PADI project, Principled Science Assessment Design for Students with Disabilities, is one of a series of projects based at SRI that seeks to address existing shortcomings in large-scale assessment in the larger framework of evidence-centered design (see Haertel, Haydel DeBarger, Villalba, Hamel, & Mitman Colker, 2010, and Hansen, Mislevy, Steinberg, Lee, & Forer, 2005, for the theoretical underpinnings of the project). This particular project investigates the value of integrating UDL considerations into the task design process. This work represents, to our knowledge, the first significant attempt to systematically apply UDL principles in conjunction with evidence-centered design principles to the design of large-scale assessment items. This report articulates the principles of UDL, emphasizing their application to assessment, and provides some examples of those principles as they are presently being applied in the PADI project.

2.0 Origins of Universal Design and Universal Design for Learning

The term “universal design” originally comes from the field of architecture where the emphasis has been on the design of products, buildings, or environments that can be used readily

by the widest possible range of people (Mace, Hardie, & Place, 1996). Virtually all architects in the U.S. now create buildings that are designed from the outset to reduce or eliminate architectural barriers through designs that consider the diverse needs of different people. This practice now is recognized as highly more cost-effective and equitable than trying to retrofit buildings later or providing customized accommodations to individuals who are unable to navigate poorly designed structures. Universally designed environments are engineered for flexibility and designed to anticipate the need for alternatives, options, and adaptations to meet the challenge of diversity. While originally conceived to meet the needs of individuals with disabilities, universal designs has proven to make buildings more accessible and functional for everyone.

A good example of universal design in action comes from the history of television captioning. When captioning first became available, it was an expensive add-on purchase intended for people with hearing impairments. Building captioning into every television, rather than retrofitting it later, turned out to be a better, and more universal, design. It now benefits not only those with hearing impairments, but also exercisers in health clubs, travelers in airports, individuals working on their language skills, and couples who go to sleep at different times. The key to universal design is building options into initial designs, thereby making better choices available to everyone.

Universal Design for Learning (UDL) is one aspect of the overall movement toward universal design. UDL places the focus on *learning* and on learning environments or curricula. For many, UDL is synonymous with providing access to information. But that view is too narrow. While providing access to information is often essential to learning, it is only one aspect of UDL. Learning typically encompasses many kinds of changes in performance and capacity that go far beyond the mere acquisition of information. Moreover, providing access to learning, as opposed to mere information, requires that the *means* for teaching and learning — the pedagogical goals, methods, materials of instruction — are themselves accessible to all students. UDL is the process by which we attempt to ensure that the means for learning, and their results, are equally accessible to all students.

For the most part, educators have sought to apply the principles of UDL to the design of instructional materials – the books, instructional technologies, and curricular materials found in classrooms. Applications of UDL to methods of teaching, the process of setting goals, and assessment, especially large scale assessment, lag behind.

2.1 The Basic Framework of UDL

The framework and guidelines for UDL are not derived from the principles for architecture, but rather from research and practice from multiple domains within the “learning sciences:” Education, developmental psychology, cognitive science, and cognitive neuroscience. The research in these fields guides both the scope of the pedagogy that UDL addresses (i.e., the critical elements of teaching and learning) and the range of the individuals that UDL addresses (i.e., the critical elements of individual differences).

At its simplest, the scope of UDL is based entirely on three principles:

Provide Multiple Means of Representation
Provide Multiple Means of Action and Expression

Provide Multiple Means of Engagement

These three principles address three critical features of any teaching and learning environment: the means by which information is presented to learners, the means by which learners are required to interact with materials and express what they know, and the means by which students are engaged in learning (for further details, see Rose & Meyer, 2002, and Rose, Meyer, & Hitchcock, 2005).

While there are many ways to articulate the fundamentals of teaching and learning, the choice of these three foundational principles stems from their commonality across many aspects of theory and research in the learning sciences. Consider the field of cognitive neuroscience, where it is common to think of three broad divisions of the “learning brain”: 1) the pattern recognition capabilities in the posterior regions of cortex, 2) the motor and executive capabilities in the frontal regions of cortex, and 3) the affective or emotional capabilities in the medial regions of the nervous system. While even this division is an over-simplification, it is an articulation that is common and draws historically on Luria’s classic work (Luria, 1973), and has been elaborated and modified by many others (e.g., Cytowic, 1996; Goldberg, 2001; Barsalou, Breazeal, & Smith, 2007; Rosenzweig, Breedlove, & Watson, 2005). In order to be systematic in considering learning differences, it is by design that the three principles of UDL match up well with this framework from neuroscience, addressing in turn the perceptual learning of the posterior cortex, the strategic and motor learning of the anterior cortex, and the affective or emotional learning of the medial and orbital frontal cortex.

Beyond cognitive neuroscience, however, researchers and theorists in other learning sciences have adopted very similar frameworks to consider the scope of teaching and learning. Among the most prominent, Lev Vygotsky (1978), the preeminent Russian psychologist, and Benjamin Bloom (1994), the American educational theorist, both adopted a similar three-part framework for their foundations.

From the three principles, a total of nine guidelines have been developed that form the primary foundation of UDL. While these guidelines articulate the three principles, their main purpose is to guide educators and curriculum developers in using evidence-based means for addressing the wide range of individual differences present in any typical classroom. (For more detail on the nine guidelines, see www.udlcenter.org.) In the past, it has been extremely difficult to provide sufficient alternatives to meet the challenges of diversity. Fixed textbooks and large classes make it difficult to individualize or accommodate individual differences. By taking advantage of the power and flexibility of modern technology, UDL provides a promising vehicle for delivering these alternatives in educational settings.

3.0 UDL and Large-Scale Assessment

In recent years, there has been considerable interest in the application of UDL principles to assessment, especially large-scale assessment. That interest stems from two sources. The first is inherent in the more general application of UDL to education. That is, only when all elements of the learning environment — the goals, the methods, the materials, and the assessments — are universally designed can UDL be implemented effectively. Assessment is one of the critical elements of instruction, and UDL assessments are thus critical to any effective instantiation of UDL.

The second source of interest in the application of UDL to assessment comes more directly from the assessment industry. In this case, the focus is on improving the accuracy and validity of large-scale assessment. The question being asked is: Can the application of UDL principles improve large-scale assessment itself? Our work with PADI focuses on this latter question, to which we now turn.

3.1 UDL and the Optimization of Validity.

A critical limit on the usefulness of any assessment is its validity – does it actually measure what it is supposed to measure? Does it measure the right thing? Traditional assessments, especially large-scale assessments, usually privilege ease of scoring and the standardization of tasks. As a result, their scope can be too narrow and too shallow. Their scope is too narrow when they fail to validly measure many kinds of learning. They are too shallow when they measure only the results of learning (performance) and not the changes in constructs, skills and strategies, and motivations that are the critical sources of that performance and more predictive of future learning.

The limits on validity stem primarily from two sources. First, many traditional assessments reflect views of intelligence and measurement no longer current with advances in the learning sciences. As a result, they focus on item-level stability and standardization rather than the validity with which those items actually measure the underlying cognitive constructs and processes that are of interest. The higher-order cognitive and executive functions, as well as new media skills, are drastically under-sampled (Madaus, Russell, & Higgins, 2009). Second, most traditional assessments reflect the limits of print. As a tool for measurement, print places severe constraints on what can be measured validly. Those limits are evident in: 1) the limited types of information that can be displayed or evaluated; 2) the limited forms of expression or problem-solving that can be assessed adequately; and 3) the limited options for sustaining motivation or engagement (typical assessments standardize the items, conditions, and external incentives, but not the resulting individual level of motivation or engagement).

UDL assessments privilege validity in two ways. First, UDL assessments are designed within a framework based in the modern learning sciences. That framework broadens the scope of what must be assessed in order to make valid inferences about learning, ensuring that assessments are comprehensive and differentiated enough to address the full range of cognitive, executive, and affective changes that underlie learning as we now know it (versus what can be easily measured). The UDL principles require assessments that not only are broader in their scope but also deeper in what they probe. UDL assessment focuses not on the surface or statistical properties of items but on the underlying constructs and competencies that are critical to current performance and to future learning.

Second, UDL optimizes validity not solely through designing better items on large scale tests but by designing better learning environments—that is, environments where assessment is routinely and continuously embedded within the learning itself. That capacity is nearly impossible in a world of print, resulting in a separation between learning and its assessment that is a continuing threat to validity. UDL capitalizes on the interactive power of modern technologies (e.g., continuous, authentic, assessment is pervasive in gaming environments) to measure not only the outcomes of learning (e.g., whether an answer to a multiple choice item is right or wrong) but the learning itself. With modern technologies (e.g., simulations, interactives), it is possible to assess not only the number of correct answers but the strategies,

procedures, knowledge structures, and misperceptions that underlie both right and wrong answers. Furthermore, the embedded options and alternatives provide an additional advantage: they make it possible to examine the differential effects of various alternatives on learning, helping to more closely diagnose individual differences and predict what kinds of options will maximize future learning. This latter intersection of UDL and assessment is of less emphasis in the present project because its explicit goal is addressing the immediate limitations of existing large-scale assessments.

3.2 UDL and the Maximization of Accuracy

Within the UDL framework, accuracy is of paramount importance. Good decisions about instruction, future learning, and accountability all require the ability to accurately measure progress for all learners.

Traditional assessments, especially again large-scale assessments, approach accuracy primarily through arguments based on the standardization of items. The content of an item, the format in which it is presented, the requirements for response, and the conditions of testing are all controlled as much as possible so that they are the same for every learner. Accuracy and comparability is achieved at a surface level, in the sense of accurately reflected what students do in a common situation. The focus is on standardizing the *item*, but not on what is being measured—the underlying *construct*. The problem is that *equivalent surface conditions* may not provide equivalent evidence about learners. 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. The focus then is thus on *equivalent evidence*, which may require different surface conditions for different learners.

Consider, for example, a test where the “relevant construct” is some fact or principle of history. A multiple-choice test designed to assess that construct is typically standardized at the item level such that all items are exactly the same for everyone. But such an approach does not lead to accurate measurement because each item imposes its own “construct-*irrelevant*” demands on the learner—e.g., visual acuity to see the item, fluent word decoding to read the text, adequate English vocabulary to comprehend the item, familiarity with the item format, and so on. Those additional demands, labeled construct-irrelevant sources of variance by Messick (1989) and sometimes called “undesirable difficulties” in assessment literature, interfere with accurate measurement because they may be markedly different for each user. For some students, the undesirable difficulties of the item are more demanding than the construct it is designed to measure. Like a microscope whose eyepiece is out of focus, the instrument actually gets in the way.

3.3 Applying UDL to Large-Scale Assessments

To improve both the validity and accuracy of large-scale assessments, it is essential to design measurement instruments or items that are flexible, varied, and adjustable enough (rather than fixed or standardized enough) to measure constructs accurately and validly. The key is to design items so that they have sufficient options and flexibility. In doing so, it is essential, however, to distinguish between what is construct relevant and construct irrelevant. It is essential to reduce “undesirable difficulties” (i.e., those that are construct irrelevant) but not to

reduce “desirable difficulties” (i.e., those that are construct relevant). That is, the goal of UDL is *not* to make assessments that are easier (e.g., by providing options that reduce the difficulty of relevant construct) but to make them more focused and accurate, largely by reducing the “undesirable difficulties” that are sources of error.

The key idea in applying UDL principles to assessment is to provide each student with a form of a task for which construct irrelevant, or undesirable, sources of difficulty are minimized for every student, so that his or her performance depends to the largest extent possible on the construct relevant demands of the task. A task may be presented in different surface forms to different students if that is what is necessary to reduce construct-irrelevant demands to each of them, but the construct relevant demands will be equivalent in each of these forms.

To approach the task of UDL systematically within the PADI project, we chose to employ, with minor adaptations, the framework of the UDL guidelines introduced in Section 2.0. Because large-scale assessments are highly constrained in their design, we have been able to focus on a reduced subset of the original nine UDL guidelines and the threats that they address. The result is six categories of potential threats to accuracy and validity. Those categories cover all three principles (representation, action and expression, and engagement) but group or combine the guidelines to some extent. The primary difference is in the affective domain, where three affective guidelines have been aggregated into one affective category. In what follows, we will examine each of these six categories, and their application, through an exemplar item. The sample items shown here are drawn from collaborative work in the project with the Kansas State Department of Education. Illustrations use screen shots from the Kansas’s web-based authoring and delivery system, the Kansas online assessment system.¹

3.3.1. Category One: Perceptual

When perception is being evaluated—for example, in an eye exam at a doctor’s office—it is essential to standardize the perceptual conditions: the same stimulus (e.g., the same chart with the same size letters on the screen) should be presented to each patient using the same manner of stimulus presentation (e.g., each patient looking from the same distance), etc. That is true because perception (visual acuity) is in this case construct relevant—i.e., the focus of the measurement. Varying the perceptual demands of the eye exam (e.g., with some patients allowed to sit much closer to the chart than others) would undermine the accuracy and validity (not to mention the utility) of the assessment.

On the other hand, were the same item used to evaluate something *other* than perception—for example, to evaluate whether a patient knew the letters of the alphabet—then maintaining the same visual conditions would actually threaten the validity of the measurement. Using an identically appearing letter chart for all patients, for example, would undermine accurate measurement for those who had low vision. For them, the visual demands of the task (demands that are construct irrelevant when letter knowledge is being measured) would be hopelessly confounded with letter knowledge. As a result, some patients could be misdiagnosed as having weak letter knowledge skills when their actual weakness is in visual perception.

To avoid such problems, it is essential to provide options in the perceptual demands of any item. In the example here, the option of presenting the letters in multiple font

¹ We are grateful to our project colleagues from Kansas, John Poggio, Cheryl Randall, Neil Kingston, David Barnes, Abel Leon, and Mary O’Brian for their assistance, their insights, and the use of these materials and the Kansas online assessment system.

sizes is a simple solution. Providing captions on video or descriptions of images are other examples of the kinds of options that help ensure that every student has equivalent access to the information in an item. Wherever an item is measuring something other than perception, like knowledge or skills, it is essential to provide perceptual options so that the item is accurate and valid for all students.

Not all the perceptual barriers are as straightforward as the eye chart example. To illustrate the application of the perceptual category in the Kansas sample, see Figures 1 and 2 below.

As shown in Figure 1, in the original version of this item, all of the information is presented in a single homogeneous paragraph where key information is hard to distinguish or find. The revision below (Figure 2) alters the visual layout but does not alter the construct-relevant demands of the text. Here, the information is separated into three parts that are spaced out on the page, making it easier for students to find the important information. Also, an illustration—an alternative to the text-only—has been added. The actual question is placed below the illustration so that it is easier for students to refer back to it. In sum, while the same information is available in both versions of the skateboarder item, the perceptual demands are quite different.

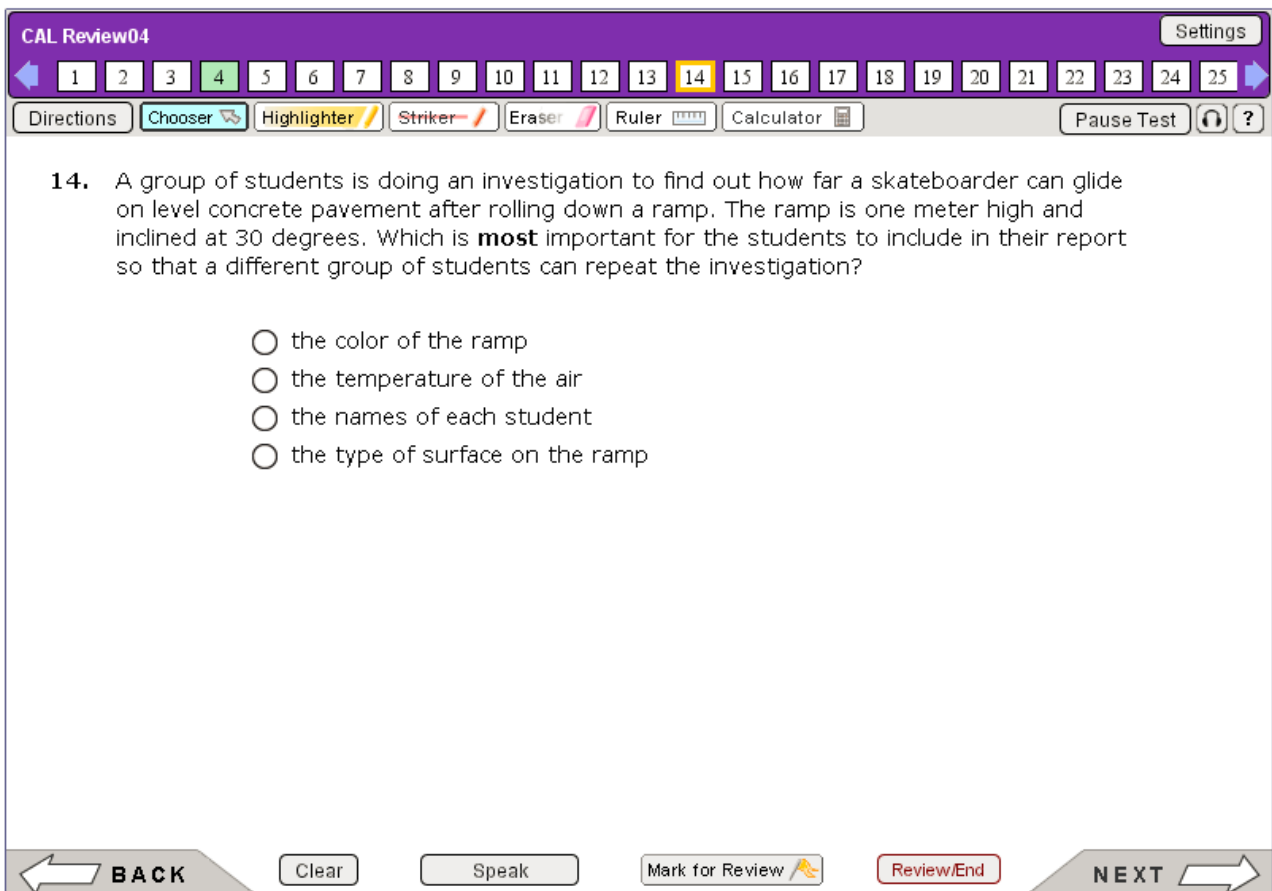
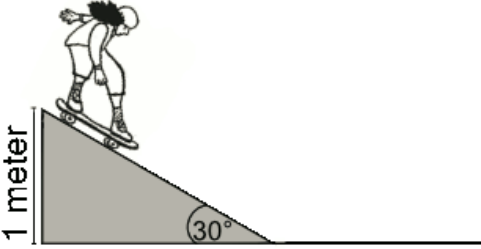


Figure 1. Skateboarder Investigation Item Original
From the Kansas Online Assessment System, Kansas State Department of Education

CAL Review04 Settings
 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
 Directions Chooser Highlighter Striker Eraser Ruler Calculator Pause Test ?

13. A group of students is doing an investigation to find out how far a skateboarder can glide on level concrete pavement after rolling down a ramp.

The ramp is one meter high and inclined at 30 degrees.



Which is **most** important for the students to include in their report so that the investigation can be repeated?

- the color of the ramp
- the temperature of the air
- the time of day
- the type of surface on the ramp

BACK Clear Speak Mark for Review Review/End NEXT

Figure 2. Skateboarder Investigation Item Revised

Adapted from the Kansas Online Assessment System, Kansas State Department of Education

3.3.2. Category 2: Language and Symbols

Much of the information in an item is not conveyed directly but instead is encoded in language and symbols. When students vary in their ability to decode those symbols (which is almost always the case), such encodings can introduce construct-irrelevant sources of error. That is, some students will face construct-irrelevant difficulties because they are not fluent in decoding, don't know the language well, or are unfamiliar with specific vocabulary. These problems, differentially distributed in the population, will interfere with accurate measurement of the intended construct for at least some students. Providing alternatives that reduce those undesirable difficulties is essential to obtain valid information about these students.

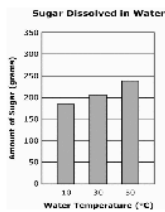
As shown in Figure 3 for an existing item about dissolving sugar, the term "milliliters" may pose a barrier for some students (and not for others). If we assume that knowledge of this term is irrelevant to the construct being measured, students' understanding of this term should not interfere with measurement.

CAL Review04

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21

Directions Chooser Highlighter Striker Eraser Ruler Calculator

4. The graph below shows the results of an investigation to find out how much sugar dissolved in 100 milliliters (mL) of water at different temperatures.



click to enlarge

Which is the **best** estimate of how much sugar will dissolve when the water is 40°C?

200 grams
 225 grams
 250 grams
 275 grams

BACK Clear Speak Mark for Review Review/End NEXT

Figure 3. Sugar Dissolved in Water Item Original
 From the Kansas Online Assessment System, Kansas State Department of Education

As seen in the revised item (Figure 4), the term “milliliters” is replaced with a less distracting abbreviation, “mL,” and this abbreviation is hyperlinked to a glossary where the correct definition is immediately provided. Such a hyperlinked glossary reduces gaps in comprehension because it fills in the gaps as needed but without the distraction and memory load of looking a word up in a dictionary.

It is worth noting that whether or not knowledge of the term “millimeters” is relevant to the construct being assessed is not a property of the item, but a property of the item in relation to the intent of the assessment. Is it part of the knowledge we want to be learning whether students have, or is it a potential barrier to our learning about the knowledge we care about? This determination that must be made through an understanding of the purpose of the assessment--for example, by reference to standards documents or instructional objectives. In fact the same demand can be construct-relevant for use of the item in an assessment and appropriate to keep in, but construct-irrelevant and to be avoided in a different use. And it may be construct-irrelevant but still appropriate to keep in if we know that all students who will be assessed are sufficiently familiar with it that it will not pose undesirable difficulty for them.

CAL Review04

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21

Directions Chooser Highlighter Striker Eraser Ruler Calculator

3. The graph below shows the results of an investigation to find out how much sugar dissolved in 100 mL of water at different temperatures.

A milliliter is a thousandth of a liter.

Water Temperature (°C)	Amount of Sugar (grams)
10	185
30	205
50	240

click to enlarge image

Based on the graph, how much sugar do you predict would dissolve when the water is 40 degrees Celsius?

200 grams
 225 grams
 250 grams
 275 grams

BACK Clear Speak Mark for Review Review/End NEXT

Click here and drag to move this window

Sugar Dissolved in Water

Amount of Sugar (grams)

Water Temperature (°C)

Close

Figure 4. Sugar Dissolved in Water Item Revised

Adapted from the Kansas Online Assessment System, Kansas State Department of Education

3.3.3. Category 3: Cognitive

Most large-scale assessment items require skills and strategies that are usually called “cognitive” and involve selective attending, integrating new information with prior knowledge, strategic categorization, active memorization, and the like. Often at least some of these are construct relevant (i.e., they are part of what is intended to be measured). But usually there are also some cognitive demands that are not relevant—that is, they are introduced by the particular way in which the item is represented and by the demands that the item’s representation places on the learner. For example, the posing of a problem for a mathematics assessment may place high-level cognitive comprehension demands on the learner. These demands are not relevant to the construct being measured but are instead imposed by the way the item is constructed. A long paragraph of explanation, for example, poses many problems for reading comprehension and memory that add difficulty to the item but may not be at all relevant to the mathematical construct being measured. If the added difficulty were the same for all students, the effect would be unimportant. But that would rarely be the case. Students inevitably vary considerably in the information processing skills they bring to the item. As a result, the item becomes unreliable as a measure of the construct.

For the item on magnets shown as Figure 5, the relevant construct pertains to understanding the properties of magnets and their attraction. In addition, however, from looking at the representation, students must be able to recognize and distinguish that there are two poles of the magnets, north and south, symbolized merely by an “n” and an “s.” For many students, this distinction would be obvious or at least relatively easy to discern. Others would struggle or fail to make this distinction and, thus, would do poorly on the conceptual question that depends on recognizing that distinction.

CAL Review04 Settings

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25

Directions Chooser Highlighter Striker Eraser Ruler Calculator Pause Test ?

8. A magnet was placed between two other magnets, as shown in the picture below.

N	S
---	---

↓

N S	S N
-----	-----

Which shows what **most likely** happened when the magnet was placed between the other two magnets?

N	S
---	---

N	S
---	---

S	N
---	---

N	S	N	S	S	N
---	---	---	---	---	---

N	S
---	---

N	S	S	N
---	---	---	---

N	S
---	---

N	S
---	---

S	N
---	---

BACK Clear Speak Mark for Review Review/End NEXT

Figure 5: Magnets Item Original
 From the Kansas Online Assessment System, Kansas State Department of Education

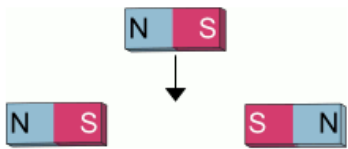
To address potential barrier in a revision (see Figure 6), the team highlighted the critical features of the magnets. The poles are now color-coded such that the north end is colored in blue and the south end is colored in red. Such highlighting draws attention to the critical features of the magnets but does not reduce the difficulty of the relevant conceptual question being asked (i.e., what will happen when the new magnet is interposed?). Such highlighting of critical features is one of the important scaffolding processes of UDL that lies within the cognitive category.

CAL Review04 Settings


1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25


Directions Chooser Highlighter Striker Eraser Ruler Calculator Pause Test ?

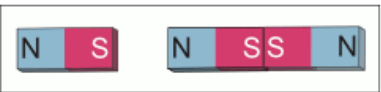
7. A magnet was placed between two other magnets, as shown in the picture below.




Which shows what **most likely** happened when the magnet was placed between the other two magnets?









BACK Clear Speak Mark for Review Review/End NEXT

Figure 6: Magnets Item Revised

Adapted from the Kansas Online Assessment System, Kansas State Department of Education

3.3.4. Category 4: Skill and Fluency

All assessment items require some kind of response from students. Often the physical demands for the response—for example, choosing one answer for a multiple choice item—are relatively trivial. At other times, the method of responding—for example, writing an essay—can be more cognitively and executively demanding than the constructs the item is intended to measure. All means of responding introduce demands on the examinee, demands that are commonly construct-irrelevant, potentially distracting, and as such, threats to validity. Even the relatively trivial action required for a multiple choice item is very difficult for some students. For students with physical disabilities, say, the effort of physically responding may be equally or even more challenging than the construct-relevant demands of the item. By providing no alternatives for responding, the physical challenge is confounded with the cognitive challenge, and the validity of the item is severely compromised for those students.

Most of the threats to validity associated with responding come not from the sheer physical demands of response but from the skills and fluencies that are implicitly required in the means of response. In writing an essay on a history item, the skills of handwriting and spelling are largely implicit. For most students, the additional demands of these skills are relatively

trivial. For students with dysgraphia or dyslexia, these demands are far from trivial. As a result, failure to provide options sharply reduces the validity of such items.

Many assessment items, like the one shown in Figure 7, require students to perform calculations. When mental calculation is the construct being measured, all students should perform those calculations without special tools. But many times the calculations themselves are not construct-relevant but only a step in determining the correct answer. For some students, even simple mental calculations are challenging, interfering with optimal performance on any item that implicitly or explicitly requires mental calculation. To eliminate this barrier, and to reduce this source of error more generally, it is important to include a calculator for all test items that might involve calculation. Note that in Kansas interface for the original item, the calculator already is provided systemically in the tool menu bar. For students whose calculations are not fluent, providing this calculator ensures that they can focus on the question content rather than putting most of their effort into performing calculations.

We note in passing that providing the calculator can reduce demand on calculations when they are construct irrelevant, but at the same time introduces demands for using the representations and affordances of the calculator tool – themselves construct-irrelevant as well. This would not pose a threat to validity if it were known that using the calculator was familiar to all students being assessed; construct-irrelevant demands are introduced, but they are within the capabilities of the students and unlikely to be a source of poor performance. This situation can be assured if students are familiar with classroom use or previous experience with the tool. The point isn't whether the tool is present or not, but tool's availability in conjunction with the students' profiles of capabilities—in this case, whether these particular students' ease with using the tool will reduce construct-irrelevant demands in the main.

CAL Review04 Settings

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25

Directions Chooser Highlighter Striker Eraser Ruler Calculator Pause Test ?

23. The diagram below shows the direction and amount of two forces acting on a box. The forces are measured in Newtons.

80 N \longrightarrow \longleftarrow 60 N

What is the **net force** acting on the box?

20 N to the right \longrightarrow 20 N

60 N to the left \longleftarrow 60 N

80 N to the right \longrightarrow 80 N

140 N to the left \longleftarrow 140 N

BACK Clear Speak Mark for Review Review/End NEXT

Figure 7: Calculation of Force Item Original

From the Kansas Online Assessment System, Kansas State Department of Education

3.3.5. Category 5: Executive Functions

Many assessment instruments place high demands on what are called “executive functions”—that is, the abilities involved in inhibiting impulsive, short-term, immediate responses in favor of those that are associated with strategic, executive thinking (e.g., careful goal-setting and planning, selection of effective strategies for reaching goals, consistent monitoring of progress, etc.). In some items, those executive functions are construct-relevant because the intent is to measure a student’s ability to plan, execute, and monitor progress over time. But in other items, executive functions are construct-irrelevant and get inadvertently inserted because of the way the item is constructed (e.g., being required to write a short essay to demonstrate a knowledge of facts in musical history, when in the assessment is measured by the historical facts rather than the strategy and planning involved in organizing the essay). Like any other ability, executive function is highly variable in any population. As a result, the imposition of executive function demands—at least those that are construct-irrelevant—can renders items invalid and inaccurate for some students all the time and for many students some of the time. For some students, the demands of executive function overwhelm the construct relevant demands of the item.

As a result, it is essential to provide options or alternatives for the executive demands of items that are not construct-relevant, for students for whom the demands would present nonignorable sources of difficulty. Many such options are described in the UDL guidelines. In an original item detailing a toy truck experiment (see Figure 8), for example, a lot of information is packed into an introductory paragraph. To unpack that information requires numerous executive functions: the ability to set up a plan for retrieving information buried in the paragraph, the ability to hold that information in working memory and organize it, the ability to implement the strategy systematically, and the ability to monitor progress (e.g., “Did I include *all* the information?”).

CAL Review04 Settings

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25

Directions Chooser Highlighter Striker Eraser Ruler Calculator Pause Test ?

2. A student wants to know how different weights placed in a toy truck affect the truck's speed down a ramp. She has gathered a toy truck, three rocks, a stopwatch, and a ramp. For each test, she will place a different rock in the truck and time how long it takes for the truck to roll down the ramp. The student will record her data in the table below.

Truck Experiment

Test	Rock Weight (grams)	Time for Truck to Roll Down Ramp (seconds)
1	200	
2	300	
3	300	

Which part of this plan could make it harder for her to answer her question?

- She is using the same ramp for each truck.
- She is using a stopwatch to record the time.
- She is using two rocks with the same weight.
- She is using only one truck for all three tests.

BACK Clear Speak Mark for Review Review/End NEXT

Figure 8: Toy Truck Experiment Item Original
From the Kansas Online Assessment System, Kansas State Department of Education

In the original item, the data table is an excellent device that scaffolds the student's executive functioning without diminishing the scientific thinking involved. The team recommended using a similar scaffold (a simple checklist) for the materials. In the revised item (see Figure 9 below), the sentence describing the various materials used in the experiment is removed from the paragraph and transformed into a bulleted list located next to the data table. This redesign supports students in managing information and resources; the materials are now listed in an organized manner. The revised item assesses the scientific concepts more directly and reduces the intrusion of executive functions that are not relevant to the science involved.

CAL Review04 Settings

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25

Directions Chooser Highlighter Striker Eraser Ruler Calculator Pause Test ?

1. A student wants to know how different weights placed in a toy truck affect the truck's speed down a ramp. Below is her list of materials and the table she will use to record her data. For each test, she will place a different rock in the truck and time how long it takes for the truck to roll down the ramp.

Truck Experiment

<p><u>Materials</u></p> <ul style="list-style-type: none"> <input type="radio"/> - toy truck <input type="radio"/> - three rocks <input type="radio"/> - stopwatch <input type="radio"/> - ramp 	<p><u>Data table</u></p> <table border="1" style="width: 100%; border-collapse: collapse; text-align: center;"> <thead> <tr> <th>Test</th> <th>Rock Weight (grams)</th> <th>Time for Truck to Roll Down Ramp (seconds)</th> </tr> </thead> <tbody> <tr> <td>1</td> <td>200</td> <td></td> </tr> <tr> <td>2</td> <td>300</td> <td></td> </tr> <tr> <td>3</td> <td>300</td> <td></td> </tr> </tbody> </table>	Test	Rock Weight (grams)	Time for Truck to Roll Down Ramp (seconds)	1	200		2	300		3	300	
Test	Rock Weight (grams)	Time for Truck to Roll Down Ramp (seconds)											
1	200												
2	300												
3	300												

Which part of this plan could make it harder for her to answer her question?

- Using the same ramp for each truck.
- Using a stopwatch to record the time.
- Using two rocks with the same weight.
- Using only one truck for all three tests.

← BACK
Clear
Speak
Mark for Review
Review/End
NEXT →

Figure 9: Toy Truck Experiment Item Revised

Adapted from the Kansas Online Assessment System, Kansas State Department of Education

3.3.6. Category 6: Affective

Most large scale assessments require sustained attention and effort. When motivated, many students can regulate their attention and affect in order to sustain the effort and concentration that assessments require. However, students differ considerably in their ability to self-regulate in this way. We would apply the “Affect” UDL category to support self-regulation. Among the methods that UDL recommends to help students sustain effort and persistence are scaffolds that make goals more explicit and visible (especially, the progress toward them). These scaffolds include modifications that help students keep track, and be rewarded by, their progress toward completion of the goals.

Fortunately, the Kansas Online Assessment System already includes an explicit scaffold that allows students to more easily see the “finish line” and to monitor their progress. As shown in the Figure 10 example item, the bar at the top of each screen lets students easily see how many items there are in total for the goal of completing the assessment and allows them to easily keep track of their progress toward that goal. In addition, at the bottom of the screen, students can mark items they are unsure of and return to them later, allowing them to navigate around specific “potholes” that may impede their overall progress and contaminate subsequent performance with

negative emotion (i.e., the frustration, distraction, or anxiety produced by a few difficult items early in an assessment). Finally, the Kansas interface has a “striker”—a tool that students can use to cross out answers that they are certain are incorrect, thereby reducing the cognitive and affective demands of too many choices and too much uncertainty.

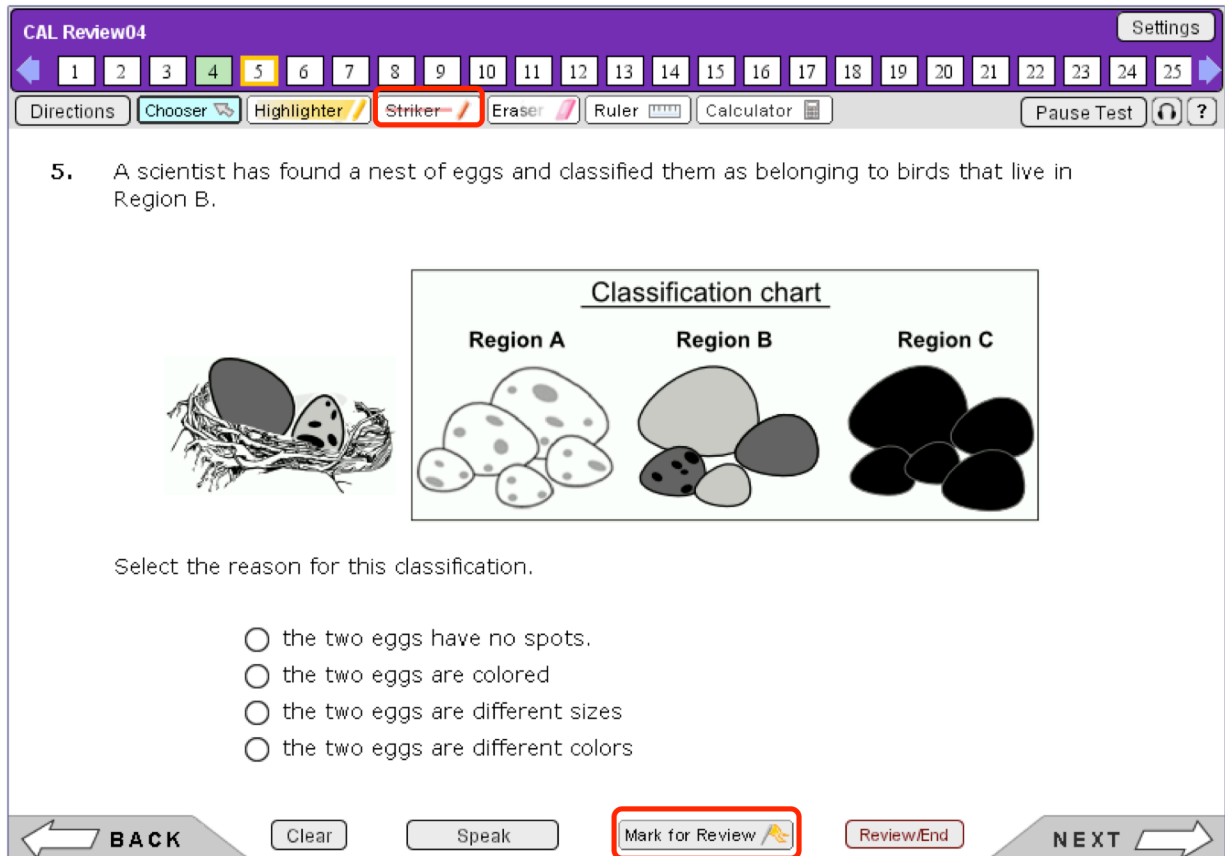


Figure 10. Bird Nest Item Original
From the Kansas Online Assessment System, Kansas State Department of Education

4.0 Conclusion

The Principled Science Assessment Design for Students with Disabilities project is exploring the intersection of the principles of the PADI process and the principles of UDL. The goal is to bring UDL principles into task design from the very start: With a clear understanding of what knowledge or skills are the target of assessment, it becomes possible to identify the construct-relevant aspects of a task idea, and consider a full range of ways that the essential challenge can be instantiated in a variety of forms that may differ with regard to features affecting interactions with the task, response production and affect. Each student should be matched with a form of the task for which construct-irrelevant demands are unlikely to pose undesirable sources of difficulty. This approach will result in assessments with stronger validity arguments, not only for students with disabilities but for all students: Alternative explanations for poor performance other than difficulty with the target knowledge or skill have been weakened up front through the principled the design process.

The present report sketches critical elements of the background for that exploration. Six categories of UDL considerations were reviewed, their relevance to assessment arguments for tasks was discussed, and adaptations of tasks provided by the Kansas State Department of Education were used to illustrate the ideas.

What lies ahead in this project are experimental trials to evaluate whether that intersection produces significant effects, and for whom. In particular, we will investigate whether modifications targeted at certain sources of construct-irrelevant difficulty will result in better performance among students who are known to have limitations in that area (e.g., reduced reading load for poor readers, reduced cognitive load for students with executive processing limitations).

References

- Barsalou, L. W., Breazeal, C., & Smith, L. B. (2007). *Cognition as coordinated non-cognition*. *Cognitive Process*, 8, 79-91.
- Bloom, B. S. (1994). *Reflections on the development and use of the taxonomy*. in Anderson, L. W. & Sosniak, L. A.(Eds.) (1994), *Bloom's Taxonomy: A Forty-Year Retrospective*. Chicago National Society for the Study of Education
- Cytowic, R. E. (1996). *The neurological side of neuropsychology*. Cambridge, MIT Press.
- Goldberg, E. (2001). *The executive brain: Frontal lobes and the civilized mind*. New York: Oxford.
- Haertel, G., Haydel DeBarger, A., Villalba, S., Hamel, L., & Mitman Colker, A. (2010). *Integration of Evidence-Centered Design and Universal Design Principles Using PADI, an Online Assessment Design System (Assessment for Students with Disabilities Technical Report 3)*. Menlo Park, CA: SRI International.
- 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.
- Mace, R. L., Hardie, G. J., & Place, J. P. (1996). *Accessible environments: Toward universal design*. Raleigh, NC: Center for Universal Design.
- Madaus, G., Russell, M., & Higgins, J. (2009). *The paradoxes of high stakes testing*. Charlotte, NC: Information Age Publishing.
- Messick, S. (1989). *Validity*. In R.L. Linn (Ed.), *Educational measurement* (3rd Ed.) (pp. 13-103). New York: American Council on Education/Macmillan.
- Rose D. H., & Meyer, A. (2002). *Teaching every student in the Digital Age: Universal design for learning*. Alexandria, VA: Association for Supervision and Curriculum Development.
- Rose, D., Meyer, A., & Hitchcock, C. (Eds.). (2005). *The universally designed classroom: Accessible curriculum and digital technologies*. Cambridge, MA: Harvard Education Press.

Rosenzweig, M. R., Breedlove, S. M., & Watson, N. V. (2005). *Biological physiology* (4th edition). Sunderland, MA: Sinauer Associates.

Vygotsky, L. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.



Sponsor

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

Prime Grantee

SRI International. *Center for Technology in Learning*

Subgrantees

ETS

CAST

