Constructing Innovative Computer-Administered Tasks and Items According to Universal Design: Illustrative Examples with Pilot Data

Kelly S. Burling, Rebecca Beck, Jennifer Jude, Elizabeth A. Murray & Robert P. Dolan
CAST

Michael Harms
Pearson Educational Measurement

Paper presented at the annual meeting of NCME, San Francisco, CA, April, 2006
Introduction

In this era of emphasis on accountability for public education, it is critical that the assessments used to measure student learning can support valid inferences regarding system success. This is especially the case for evaluating the learning of students with disabilities, many of whom had in the past either been excluded from test design considerations or from testing altogether. Universal Design for Learning (UDL) (Rose & Meyer, 2002) proposes that access to the curriculum for diverse students can be achieved by providing multiple access points to instructional materials, instructional strategies, and knowledge demonstration. The companion paper Constructing Innovative Computer-Administered Tasks and Items According to Universal Design: Establishing Guidelines for Test Developers further suggest that by applying UDL to computer-based testing (CBT), and in particular item development, parallels between the processes involved in learning and the processes involved in understanding and responding to test items become apparent. A key issue in scrutinizing these parallels is the identification of embedded supports and access points not only appropriate during instruction but that respect the intended constructs being assessed and hence don’t invalidate inferences that can be drawn from test results (Dolan & Hall, 2001).

UDL is predicated on individual learning differences, a concept supported by recent understandings of brain function. Learning can be thought of as occurring roughly across three sets of networks in the brain: recognition networks, strategic networks, and affective networks. Recognition networks receive and analyze information, the “what” of learning. Strategic networks plan and execute, the “how” of learning. Finally, affective networks evaluate and set priorities, the “why” of learning. These networks and the pathways that connect differ across individuals such that variability in strength and weakness across and within these brain networks contributes to learning differences. Successful teaching results in students learning to recognize essential cues and patterns, mastering skillful strategies for action, and engaging with learning. Because of students’ individual strengths and weaknesses, it is important that teaching materials and strategies be highly flexible to minimize potential barriers to learning. This is the basis of UDL, for which three principles are proposed to maximize learning and minimize barriers based upon the three brain networks, as shown in Table 1.
Table 1. Principles of the UDL framework

<table>
<thead>
<tr>
<th>Principle 1: Multiple Means of Representation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Since there are no formats or media for presenting information that are equally accessible to all students, use multiple means of representation</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Principle 2: Multiple Means of Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Since there are no response formats or media through which all students can equally express what they know, use multiple means for expression</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Principle 3: Multiple Means of Engagement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Since there are no means of engagement that are equally motivating for all students, use multiple means of engagement.</td>
</tr>
</tbody>
</table>

The way in which material is presented has an impact on how the recognition network receives and begins to analyze information. If students are provided flexibility in how they can interact with and express their knowledge and skills, each student can use his or her strengths in planning and executing that expression, impacting the strategic network. Finally, when students are provided multiple options for engagement, they can choose the option that corresponds to their own priorities and interests as governed by their affective network. However, a student may prefer, or be more capable of, a certain means of expression because of their strengths across all three brain networks. Similarly, a particular representation might be more accessible to a student because he likes the subject matter rather than because he has a facility with the media. For example, a student who is not a fluent reader may prefer to learn about plot development by reading a story about a baseball player than by watching a play about a medieval royal family. This student might have an easier time comprehending spoken words, but her attraction to baseball and disinterest in medieval court life make the baseball text more compelling and a better learning tool. To suggest that a particular means of expression is linked to a particular type of strategic processing, or that a certain representation is effective only because of a recognition processing, ignores the interrelationships between the networks (Rose, Meyer, & Hitchcock, 2005). The intent of the principles is to ensure that options, or points of access, are available to support learning from the perspective of each of the three networks.

**Flexibility, Options, and Construct-relevance**

Translating the UDL principles and three networks concept to assessment requires a more detailed understanding of the ways in which student differences manifest to ensure that modifications to test items don’t change what is being measured. The objective of both instruction-based UDL and traditional test accommodations is to remove barriers between an individual and an environment; UDL as implemented in the classroom through supports and scaffolds is designed to minimize barriers on learning in the same way test accommodations are designed to minimize barriers during testing. Both are predicated on the assumption that certain barriers, irrelevant to the knowledge or skills being conveyed or measured, can negatively impact
the validity of the learning or interferences that can be drawn from test results. While during learning the media used for recognition and expression are often irrelevant (provided learning is indeed occurring), the situation is more limited during testing. The rationale for the use of testing accommodations relies on a clear definition of what a test intends to measure, which then determines those aspects of test administration that can be modified without affected the measurement of intended test constructs. Full understanding of which aspects of individual items are construct-relevant versus construct-irrelevant requires a solid understanding of how student differences determine the ways in which students perceive, interact, and respond to assessment items. This in turn means defining the factors that influence whether a student answers an item correctly or incorrectly. We refer to these factors as sources of variance. At the item level, construct-relevant variance differentiates between students’ knowledge and skills while construct-irrelevant variance (CIV) results from reliably measuring one or more sources of variance that are extraneous to the interpreted construct (Messick, 1994). This “excess” reliable variance results in invalid and imprecise assessments because it consistently depresses or inflates an individual’s score while be unrelated to the focus of the assessment. By identifying and defining sources of variance assessment developers can consciously identify those that are construct-relevant from those that are construct-irrelevant. The result is a more universally designed assessment.

The flexibility inherent in digital technologies can potentially help reduce CIV (Dolan & Hall, 2001; Dolan, Hall, Banerjee, Chun, & Strangman, 2005). Digital technologies can incorporate multiple representations (text, video, audio), transform within and across media, and incorporate such as like highlighters, linked dictionaries, and scaffolding (Rose et al., 2005). However, CBT can introduce new CIV through the use of novel media, item types, and interfaces. Universal design for computer-based testing (UD-CBT) establishes a framework for helping test developers determine which solutions to implement and when in order to (1) ameliorate source of CIV that would otherwise be present in traditional testing, (2) identify the sources of variance that digital media and environments introduce to testing, and (3) provide design solutions for all identifiable sources of variance.

**Universal Design and Computer-based Testing: Principles and Components**

UD-CBT builds upon the central tenet of providing flexibility for students by identifying sources of variance and corresponding options that can reduce or remove the impact of the variance on student performance. These sources of variance and design options fall into the following six broad categories:

1. Perceptual
2. Linguistic
3. Cognitive
4. Motoric
5. Executive
6. Affective

As described in the companion paper, UD-CBT proposes options within each of these categories appropriate to each assessment item. An assessment item is often considered as an “atomic” entity. However, decomposing items into constituent parts simplifies the process of identifying sources of variance. Instead of attempting an exhaustive list of all possible item types and
subjecting each to an analysis of its sources of variance, it is easier to identify the individual components that make up items and the sources of variance for each such component. The following eleven item components have been identified in traditional and novel assessment items:

1. **Text**
   All items require comprehension of the instructions – students must be able to understand what is being asked of them in order to respond. This is generally accomplished using text. This level of comprehension should always be supported. Construct-relevant text in the instructions and stimulus materials, however, may or not supported, as such supports may interfere with intended constructs, as in the case of items assessing text decoding skills on a English language arts exam.

2. **Images**
   Photos, static images, icons, symbols

3. **Audio**
   An audio recording or audio track accompanying video or animation

4. **Tables and graphs**
   Tables used to organize information, convey structure and relationships, graphs used to represent data visually

5. **Mathematical numbers and symbols**
   Mathematical notation

6. **Video and animation**
   Dynamic photos and images

7. **Response options**
   From selected response to complex response

8. **Active Objects/Links**
   Drag and drop objects, radio buttons, objects whose appearance changes as a result of student input or action

9. **Multi-stage / multi-part items**
   Items that consist of multiple steps

10. **Constructed response – text**
    From fill in the blank to essays

11. **Constructed response – math**
    From a single number, to complex proofs or displays of work

As new media or item designs are utilized, they can be added to this list and to the UD-CBT framework. The goal of the current study was to evaluate the components in this list was analyzed for sources of variance in each of the six UD-CBT categories, and suggest design options for each source of variance.

**Methodology for Evaluating UD-CBT Design Considerations**

Students with disabilities and their teachers were interviewed in one-on-one sessions. The basis of these interviews was a series of “item concepts” – prototype assessment items that contained new media and formats. These item concepts were developed based upon the principles of UD-

---

*This component does not yet have a table. We hope to learn more about the issues associated with understanding and navigating multi-stage and multi-part items during our student and teacher interviews.*
CBT and on feedback on design options and the impact of new media in assessments solicited from experts in the fields of educational assessment, disabilities education, and assistive technology. A total of 35 UD-CBT test item concepts were developed, covering the following four subject areas (reading/language arts, mathematics, science, and social studies) and three grade spans (grades 3-5, grades 6-8, and grades 9-12). Each item concept consisted of a brief explanation of how the item would look and function, a list of the salient design features, and a storyboard and/or individual images or screenshots, and was accompanied by a clear articulation of the intended construct from a state’s curriculum or a national organization. All item concepts were analyzed with the UD-CBT design tool to identify the aspects of the media components and item design that introduce sources of construct-relevant and construct-irrelevant variance. The tool also presented design options to alleviate sources of CIV that are potential threats to precision and validity. Many of the design options were represented in the item concepts, and those that were feasible were implemented using storyboards rather than single static images for the interviews.

Time-constraints and student attention were primary considerations in designing and selecting the items to use in the interviews. Ten of the original 35 UD-CBT test item concepts were selected for use in the interviews based on containing media and tools that would generate the most relevant student feedback. The items were also chosen to ensure that students from all grade ranges had the opportunity to interact with as many media components as possible. Table 2 presents the item concepts and their associated media and tools identified as target features chosen for the interviews. Screenshots of the item concepts are provided in Appendix A. Item Concepts.
Table 2. Test item concepts and target features

<table>
<thead>
<tr>
<th>Name of Item Concept</th>
<th>Grade Range and Subject</th>
<th>Target Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Caribou</td>
<td>3-5 / Reading</td>
<td>Graphic organizer, Text-to-speech, Highlighting, Reading selection broken into sections with questions</td>
</tr>
<tr>
<td>2. Coins Make A Dollar</td>
<td>3-5 / Math</td>
<td>Images, Drag and drop</td>
</tr>
<tr>
<td>3. Floating Rock</td>
<td>3-5 / Science</td>
<td>Animation with accompanying text</td>
</tr>
<tr>
<td>4. Bacterial Reproduction</td>
<td>6-8 / Science</td>
<td>Animation with accompanying text</td>
</tr>
<tr>
<td>5. Fraction Finder</td>
<td>6-8 / Math</td>
<td>Multi-stage, Complex instructions, Revisiting previous stages</td>
</tr>
<tr>
<td>6. Geological Timeline</td>
<td>6-8 / Science</td>
<td>Multi-step, Drag and drop, Clarity of instructions</td>
</tr>
<tr>
<td>7. IKEA</td>
<td>6-8 / Civics</td>
<td>Video with long verbal description, Composition tools</td>
</tr>
<tr>
<td>8. NAFTA</td>
<td>9-12 / Geography</td>
<td>Image based item, Long descriptions</td>
</tr>
<tr>
<td>9. Principles of Gasses</td>
<td>9-12 / Math</td>
<td>Animation with video</td>
</tr>
<tr>
<td>10. Ratio of Inscribed Figures</td>
<td>9-12 / Math</td>
<td>Drag and drop, Images, Enabled elements</td>
</tr>
</tbody>
</table>

**Interview Protocol**

Interviews were designed to last 45 to 60 minutes for students and 30 to 40 minutes for teachers. Interview questions were developed with the following concerns in mind. If the questions were initially targeted toward specific media and tools, they could curtail students’ opportunities to share richer observations that may contribute to the findings in unexpected ways; by directing students’ attention to the features we thought salient, we might miss the opportunity to learn what drew their attention or concern. For this reason a broad focus was adopted for the initial questions, followed by more specific item design concepts targeting as the interview progressed, with the goal of providing students with ample opportunities to elaborate and thus encourage discussion of specific media and tools. In this way the protocol most closely resembled a structured, retrospective think-aloud or cognitive lab.
The interview sequence for each item concept typically began with a prompt to the student to read through the item, explore its features, and treat it as if it were a real test question, followed by a series of open-ended questions, as follows:

1. Query regarding item task (“What do you think this items is asking you to do?”).
2. Supported, screen-by-screen analysis of multi-stage items, if students couldn’t define the task initially.
3. Targeted questions about elements of the media and/or tools (and system level tools if they were relevant to the item).
4. Questions about the relative ease and usefulness of the test item compared to traditional tests and why students held the opinions they did.
5. Questions related to engagement of the test item.
6. Targeted questions related to the student’s specific disabilities and whether the test item reduced any barriers that typically exist during traditional testing.

The rationale for the initial broad focus was that if students can accurately describe an item’s intent after an open-ended exploration, then the media and tools must be clear and intuitive enough to create transparency. If the media and tools were confusing, students would have a difficult time ascertaining the intent of the item, thus requiring more prompting and support from the interviewer. When practical, a second, reworded version of a question was asked of students to verify their responses. For example, a question such as “Did the use of images make it easier for you to organize your thinking and plan a response?” may have been followed with “Do you feel the images gave you enough information for writing a paragraph on the impact of NAFTA? Why or why not?”

In order to provide quantifiable measures of the student responses, a five-point Likert scale was used for discreet questions that were amenable to numeric ratings, such as those relating to an evaluation of clarity, confusion, facility of use, and usefulness of specific tools for the test item.

In addition to following an interview sequence that progressed from broad questions to more targeted ones, the questions were also organized around descriptive terms that would allow for a more cohesive analysis of student responses. Three main descriptors were adopted as main categories for describing the items and for organizing student responses: overall clarity, facility of use, and engagement.

**Overall Clarity**

An item was considered to have overall clarity if it offered a clear presentation of the intended construct and its media and tools. If students were able to accurately describe what was expected of them with little or no prompting, then the item was rated as having high overall clarity.

**Facility of Use**

This descriptor encompassed student responses that pertain to the ease with which students could interact and respond to construct-irrelevant aspects of the item (i.e. ignoring construct-relevant challenges).

**Engagement**

Items rated as engaging were those that maintained high levels of student interest due to the aid of specific UD-CBT media, item tools, and features.
Selecting Students: Considering the Margins

UDL is focused on the notion that studying the needs of students “in the margins”, from students who struggle to those who excel, will inform curricular and assessment solutions for the entire spectrum of learners (Rose et al., 2005). Just as a majority of individuals benefit from curb-cuts — originally designed for wheelchair users — when we’re pushing a baby stroller or wheeling a suitcase, so do have all learners been found to benefit from the supports built into digital learning and assessment media for individuals in the margins. For example, when visual fatigue sets in for an otherwise skilled reader, a text-reader can read an article from the web; voice activation can provide short term assistance to someone who is temporarily unable to use the keyboard readily, perhaps due to Carpal Tunnel Syndrome or simply simultaneous engagement in other manual tasks. The UD-CBT project extends this idea of learning from the margins to assessment situations. This interview process targeted students with disabilities, yet in the long run it is expected that what we learn from these students may help all students.

Students interviewed all participated in the regular state achievement tests, and were in either integrated general education or special education classrooms. The most common testing accommodations used by the students in the study included read-aloud, extended time, calculators, small group settings, and in two cases, the use of graphic organizers. Table 3 outlines the numbers of students for each of the disability categories and grade levels. A total of 33 students were interviewed.

Table 3. Participating students

<table>
<thead>
<tr>
<th>Exemplar Categories by Grade</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blind/Low Vision</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Deaf/Hearing Impaired</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reading Disability/Dyslexia</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Math Disability</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cognitive Disability</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asperger’s Syndrome</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Attention Deficit/Hyperactivity Disorder</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multiple Disabilities</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4</td>
</tr>
</tbody>
</table>

If students’ disabilities crossed more than one category, they are noted in the multiple categories column. There was one student with multiple disabilities from each of the following categories: hearing impaired & cognitive disability (grade 3), reading disability & Asperger’s (grade 4), hearing impaired and reading disability (grade 5), attention deficit/hyperactivity disorder & reading disability & cognitive disability (grade 12), reading disability & social-emotional (grade 12), attention deficit/hyperactivity disorder & reading disability (grade 12), and reading disability & math disability (grade 12).

In addition the students, a total of eight teachers and one vision specialist, all of whom worked with the students included in the study, were interviewed.
Results

In general, students and teachers were able to understand the item concepts and provide feedback as if they were actual, fully-developed assessment items. For example, the “Coins Make A Dollar” item concept presented a single-screen, single-step test item, in which students were directed to drag any combination of coins into the box to form $1.00. Even though the item wasn’t fully operational, its intent was clear to all the students interviewed. Another example, the “NAFTA” item concept, required navigation through multiple screens and reading text descriptions of images, yet 75% of the students who examined the item were still able to grasp its intent; this was true even though other factors such as lack of background knowledge of NAFTA itself would have prohibited many from actually responding correctly. In this case, even though students were unfamiliar with the content they were still able to comprehend the function of the tools and media and the intent of the item.

Results from the student and teacher interviews are presented in two ways. First, summaries of major findings are presented according to the six UD-CBT categories grouped as perceptual and linguistic options, cognitive and motoric options, and executive and affective options. Second, an analysis of student responses by item is presented by ranking items in terms of their perceived ratings for the three main descriptors: overall clarity, facility of use, and engagement.

Perceptual and Linguistic Options – Major Findings

Provide options for representing item stimuli

The majority of students indicated that when an item stimulus is presented in more than one way (e.g. animation with narration), content was conveyed more succinctly, clearly, and dynamically. This in turn appeared to increase the likelihood that the student would comprehend greater depth and detail in the item’s stimulus.

Alternatives to text were often preferred by students, and especially by those with diagnosed reading disabilities. Even students reading on grade level expressed that animation, video, and images offered more cohesive and comprehensible content than text alone. When asked if specific elements of an item helped maintained their engagement, students responses included the presence of dynamic media, such as animation, as motivation for completing an item.

Utilize multiple simultaneous media in presenting item stimuli

Students expressed a preference for the use of multiple simultaneous media in an item stimulus. They seemed to intuitively understand that with a greater variety of media, the better their chances of comprehending, and that a secondary method of presentation can back them up if the first fails. This supports one of the central tenets of UDL, namely that student performance is maximized when item design incorporates options that accommodate the greatest variety of brain functions—perceptual, linguistic, motoric, cognitive, executive, and affective. For example, students and teachers requested text-to-speech to read the text that accompanied the animation in the “Floating Rock” item and the availability of an online dictionary to present definitions of construct-irrelevant terms in the directions. Similarly, with the “IKEA” item, one student suggested offering closed captioning for the voices in the video—anther ‘safety net’ when the volume, background noise, or accent of the speaker interferes with the message.
Cognitive and Motoric Options – Major Findings

Providing background knowledge
Providing background knowledge is a key cognitive-support strategy that can help students overcome barriers due to limited construct-irrelevant prior knowledge. For instance, the intent of the “NAFTA” item was not to measure students’ knowledge of NAFTA itself but rather to measure students’ skill in evaluating a variety of perspectives and to form a viewpoint supported by available evidence. According to both students and teachers, building students’ background knowledge of the meaning of NAFTA in advance of the item would not only have helped them access the construct-relevant skills and knowledge but would have helped sustain their engagement as well.

Highlighting critical features
Based on student responses another important cognitive option, highlighting critical features, needs to be a major design consideration. One particularly astute student commented that the titles accompanying the videos in the “Floating Rock” item piqued her interest in viewing the videos and gave her some information regarding what to expect. While not all students will rely on titles and labels as an indicator of critical features, let alone even attend to them, for many this will be an opportunity for contextualizing the task and the relevance of stimuli.

Provide options for expressing knowledge and skills
In general, both teachers and students were enthusiastic about having alternatives to constructed response (i.e. prose writing) as a means for expressing knowledge and skills. In particular, the idea of an audio recording tool for composing responses was especially popular with students, but received mixed reviews from teachers.

While the options presented for expressing knowledge in this project were limited, students and teachers were queried regarding the use of non-traditional digital tools (sound, images, and drawing tools, etc.) for open-response items. This idea was well-received by both students and teachers. One teacher for instance described how critical drawing is for students with dyslexia, and she shared a process for storyboarding that her students use as a planning tool for composing longer responses. She suggested that digital stamps of common images could be made available in place of digital drawing tools in order to save time during a test. This merits consideration given what is common knowledge regarding the important role that planning plays in the construction of open responses.

Affective and Executive Options – Major Findings

Provide options for engaging with the item
The “Caribou” item is a traditional reading assessment item with the addition of tools that provide greater support for students’ strategic processing. Students responded favorably to the graphic organizer and highlighter as options to target and organize the information they needed for formulating an extended response. These simple approaches offer flexibility in the way students engage with the item. They also responded positively to the idea that the item could have two ‘view’ options—one for reading the entire text first then answering the set of questions, and another that enabled them to read the text in sections, and answer related questions as they
read the content. This latter option of course may have construct-relevant implications, but for students with disabilities may mean the difference between overall success and failure on the entire set of items associated with the stimulus.

**Color, layout, and design offer visual, sensory, and cognitive supports that help sustain engagement**

Feedback across multiple items suggests that students react favorably to color, layout, and design features that are interesting and engaging, and commented that these features helped sustain their engagement. For blind and low-vision students, layout features and color-coding (for low-vision students) are essential for negotiating their way through an item and to reducing frustration that can result in disengagement. Also important are options to clicking and dragging of interface elements for some students with sensory and motor disabilities.

**An Analysis of Student Responses By Item**

In order to summarize findings across items, the questions posed to students and teachers were assigned to three main descriptors: overall clarity, facility of use, and engagement. When more than one question was posed within one of the three categories, student responses were averaged in order to generate a single score for each category. If questions falling into one category utilized two different measures (Likert and percentage of yes/no responses), the Likert ratings were converted to yes/no responses as follows: No = ratings 1.0-2.5; Yes = ratings 3.5-5.0. A score of ‘3’ was considered ‘neutral’, and not counted in the yes/no responses.

Next, items were rank ordered within each of the three categories, permitting within-category, across-category, and across-grade-range comparisons. The rank order of items within grade ranges provides immediate comparisons of items within and across the three rating categories. In particular, comparisons of the design elements present in the highest and lowest rated items within each category supports the development of inferences regarding the appropriateness of the media, tools, and other design options contained in those items. Table 4 shows the rankings.
Table 4. Rank ordering of items

<table>
<thead>
<tr>
<th>Grades</th>
<th>Rank</th>
<th>Clarity</th>
<th>Facility of Use</th>
<th>Engagement</th>
</tr>
</thead>
<tbody>
<tr>
<td>3-5</td>
<td>1</td>
<td>Coins Make a Dollar</td>
<td>Floating Rock</td>
<td>Floating Rock</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Floating Rock</td>
<td>Coins Make a Dollar</td>
<td>Caribou</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Caribou</td>
<td>Caribou</td>
<td>Coins Make A Dollar</td>
</tr>
<tr>
<td>6-8</td>
<td>1</td>
<td>Fraction Finder / Geo Timeline</td>
<td>Geo Timeline</td>
<td>Ikea</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>IKEA</td>
<td>Fraction Finder</td>
<td>Fraction Finder / Geo Timeline</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>—</td>
<td>IKEA</td>
<td>—</td>
</tr>
<tr>
<td>9-12</td>
<td>1</td>
<td>Law of Gases</td>
<td>Law of Gases</td>
<td>Ratio Inscribed Fig’s</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>NAFTA</td>
<td>Bacterial Reproduct’n</td>
<td>NAFTA</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>IKEA</td>
<td>IKEA</td>
<td>Law of Gases / Bacterial Reproduct’n</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>Ratio Inscribed Fig’s</td>
<td>Ratio Inscribed Fig’s</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>Bacterial Reproduct’n</td>
<td>NAFTA</td>
<td>—</td>
</tr>
</tbody>
</table>

As an example, consider the “Floating Rock” item. Its high ranking in terms of facility of use and engagement agree with the fact that the majority of students were intrigued by the animation and graphics and were aware of how the animation bypassed a lengthy text explanation that would have been present in a traditional test. However, it was less highly ranked for its overall clarity. This was possibly due to content that was unfamiliar to the students, the novelty of the media, the complexity of the steps involved in accessing the media and responding, or that the text accompanying the animation was too formal and dense for this grade range.

Across all the grade ranges, the items ranked most engaging were those that used media components in ways novel to assessment. In grades 3-5 “Floating Rock” was the item most highly ranked for engagement. It is the only item for this grade range that contained animation or video, and it utilized images to a greater extent than either of the other items. Similarly, the only item in the 6-8 grade range that incorporated video or animation was most highly ranked. In grades 9-12, however, the video and animation items fell to the bottom of the engagement rankings. A possible explanation is that the video and animation items were traditional in that they required a multiple-choice response to a question about the action in the animations. The most engaging items for this group were those that used novel media in non-traditional ways. The “Ratio of Inscribed Figures” item required students to interact with dynamic images and to
construct a response based on their interactions. The second highest engagement rating in grades 9-12 went to the “NAFTA” item which required students to analyze and use photos in constructing their response. All of the most engaging items presented new media and generally asked students to interact with those media in novel ways.

Conversely, the items that were most highly ranked in clarity were the most traditional. They all included new media or interaction, but of the items presented they most closely corresponded to how the same material would be assessed in a traditional test. This corresponds to findings in cognitive psychology which suggest that students are more apt to understand what is being asked of them when familiar or high frequency terminology is used and when the context is explicit (Leighton & Gokiert, 2005). If familiarity leads to comprehension, it follows that the more familiar item formats would be clearer to students.

In addition to ranking the items in order to compare their features for design implications, the rank ordering also reveals information about student thinking and motivation when evaluating the items. The “Floating Rock” item was ranked high in engagement, yet not as high for overall clarity. Similarly, the “IKEA” item was ranked high for engagement, but lowest for overall clarity and facility of use. Students were adept at distinguishing between clarity and engagement, they weren’t necessarily so influenced by their enthusiasm for an item that they allowed it to effect their ratings of other attributes, such as assigning it a higher rating for overall clarity. The disjuncture between clarity and engagement points to an extremely important consideration for CBT and item development.

Both engagement and task comprehension are composed of significant sources of variance, and to some extent they also influence one another. An item could be initially engaging because of a novel media component. But if the intent of the item or the context of the question being asked is unclear, frustration can dampen engagement. Likewise, an readily comprehended item can inspire engagement because students can immediately connect to the item intent. Inferences about student performance are only valid if students understand what is being asked, and if item components do not interfere with perception, strategic processing, or responding. Thus not only are test developers responsible for removing CIV by clearly addressing the intended construct of each item and providing access to that content and the media used to express it, they must also be aware of student’s familiarity with media and tools they use in item design.

The most significant positive findings of the student interviews are students’ familiarity and facility with computers and the assortment of tools included in the test environment, and students’ engagement with the novel media and item types. The UD-CBT framework provides developers with the tools necessary to identify the sources of variance introduced by the computer interface, the tools and test and environment, and to the media components. Once the construct-relevance of each has been determined, the framework also suggests design options to remove the impact of those sources of variance that are construct irrelevant. What the framework does not and cannot do is ensure that the tools, item types, and media components will be familiar enough that they “will not thwart students’ thinking and problem solving,” (Leighton & Gokiert, 2005{NOTE:p.2}). Practice tests and pre-test instruction in tools and features can help (Poggio, Glasnapp, Yang, & Poggio, 2005), but the greatest impact will be from a stronger alignment between instructional and assessment tools and methods.
Conclusions

One of the most interesting results from the student and teacher interviews is that students overall seemed to quickly grasp the benefits of CBT. For the most part they expressed keen awareness of the challenges they personally face with everyday learning tasks, and with little prompting most of them were able to describe how the media and tools in the test item concepts would help overcome many of the barriers that currently exist for them with traditional tests. Indeed, they responded with great enthusiasm when they observed long-existing barriers being eliminated before their eyes in the UD-CBT items.

The student interview project was designed to isolate specific media and tools in individual test item concepts and elicit student and teacher feedback. By design, there was little overlap of media and tools across items within the grade ranges in order to maximize the number of different design features exposed to each student or teacher within a limited time period. While many of the students had distinct and individual responses to the item concepts, several patterns did emerge in how students reacted to media and tools. Student responses and suggestions support the UD-CBT categories and the principle of providing appropriate design options. The novelty of multimedia and the UD-CBT options that made items containing media accessible worked dynamically and synergistically to overcome some barriers that can prevent students from accurately expressing their knowledge in a testing environment. The greater the variety of representation of item stimuli and flexibility in offering options for crafting a response, the higher the likelihood that more students will have access and more accurately express their knowledge and skills. The students affirmed this when they responded positively to the flexibility of the UD-CBT items and even requested additional flexibility when offering feedback. They readily went beyond what we had conceptualized and offered fresh ideas and innovative suggestions.
References


Appendix A. Item Concepts

1. Caribou

Grade Range and Subject: Grades 3-5, Reading

Target Features: Graphic organizer, Text-to-speech, Highlighting, Reading selection broken into sections with questions

Description: This item is a digital version of a paper and pencil test item that assesses reading comprehension. Both multiple choice and open-response questions are presented. A nonfiction reading passage about Caribou serves as the item stimulus. Pop-up boxes provide supports for helping students use a graphic organizer to track the details needed for the last open-response question, a summary. A highlighter offers an optional way for students to note details.

Screenshot:

Directions:

Read the following paragraphs about how caribou survive in the mountains of northern Alaska. You’ll come across questions to answer throughout the reading. If you choose, you can read the entire selection first, then go back and answer the questions, or you can answer them during your reading. You are allowed to reread sections as needed in order to answer the questions.

At the end of this piece, you will be asked to write or record a summary about the challenges caribou face in finding food and how they are successful. Click on the graphic organizer button below to track details while you read, in order to prepare for your summary. Or, you can use the highlighting tool to keep track of important details.

A Caribou Journey
By Debbie S. Miller
Little, Brown, and Company, 1994

If you choose to, use this graphic organizer to track details while you read.
2. *Coins Make A Dollar*

**Grade Range and Subject:** Grades 3-5, Math

**Target Features:** Images, Drag and drop

**Description:** Students drag any combination of coins from a selection into a box to make a total of one dollar.

**Screenshot:**

---

Select and move coins into the box to make $1.
3. **Floating Rock**

**Grade Range and Subject:** Grades 3-5, Science  
**Target Features:** Animation with accompanying text  
**Description:** This multiple-choice item utilizes animation to both build background for the test question, and to demonstrate the processes of rock formation. Students first click on an animation in the introduction in order to view a floating rock. They are asked how such a rock could have formed. Then, they click on the animations of the three main geologic processes involved in forming rocks. Each animation and accompanying text describes a rock-formation process—igneous, metamorphic, or sedimentary. Students may view all three animations as many times as needed before selecting the one that best describes how the floating rock was formed.

**Screenshot:**

Select the picture below that correctly demonstrates how the floating rock was formed.  
Click the **finish** button when you have made your final selection.

![Floating Rock Animation Screenshot](http://www.childrensmuseum.org/geomysteries/floatingrock/a1.html)
4. Bacterial Reproduction

Grade Range and Subject: Grades 6-8, Science

Target Features: Animation with accompanying text

Description: Students are presented with a stepped animation of a reproduction process and are asked to identify whether it shows sexual or asexual reproduction, and then are required to provide an explanation why.

Screenshot:

View the animation below:

There are 4 parts to this animation:

• Press the continue button to view the first part of the animation.
• You can then press the replay button to replay the segment or press continue to view the next segment.
• At the end, press replay from the beginning if you would like to start over.

After viewing the animation, go to the next slide to answer the test question.
5. Fraction Finder

Grade Range and Subject: Grades 6-8, Math

Target Features: Multi-stage, Complex instructions, Revisiting previous stages

Description: Students are given two fractions located on a number line, with squares above them. By clicking on ‘row’ and ‘column’ buttons, they divide the squares into sections and highlight sections to represent the fraction on the number line. When they have completed these fractions, they are given a third square in which to create a representation of a fraction with a value between the two given fractions.

Screenshot:
6. Geological Timeline

Grade Range and Subject: Grades 6-8, Science

Target Features: Multi-step, Drag and drop, Clarity of instructions

Description: A partially completed geological timeline is presented to the student, together with a list of the major periods in Earth’s history and graphics of different fossils. Students drag and drop the period labels and fossils onto the timeline to indicate the correct order of periods and eras.

Screenshot:
7. IKEA

Grade Range and Subject: Grades 6-8, Civics

Target Features: Video with long verbal description, Composition tools

Description: A video of a news story from Greater Boston about an IKEA store is presented to students. After viewing the video, students are asked to compose a letter to the Somerville mayor about why IKEA should or should not be built in Assembly Square, using arguments or evidence from the news story to support their stance.

Screenshot:

http://greaterboston.tv/features/gb_20051107_ikea.html
8. NAFTA

Grade Range and Subject: Grades 9-12, Geography

Target Features: Image based item, Long descriptions

Description: Students are presented with a photo and caption describing a negative impact NAFTA. They are directed to view photos and captions of other outcomes of NAFTA, and then to choose two that represent another impact. Finally, the student composes a rationale for his or her choice.

Screenshot:

This photo and its caption represent one perspective on the impact of NAFTA. On the following pages you will see additional photos with captions. Scroll through all photos to select three to four possible choices that show a different perspective on NAFTA. Read the captions for this set, then make your final selection.

When you click on your final choice, you will be taken to a screen with your selected photo, its caption, and a space for you to enter text. You will be asked to describe how the photo you chose shows a different impact of NAFTA. Your response should be approximately one paragraph.

Click Next to move to the photos.
9. Principles of Gasses

Grade Range and Subject: Grades 9-12, Math

Target Features: Animation with video

Description: Students are presented with a voiced-over animation that describes one of the principles of gasses, then answer a multiple choice question: “Choose the law that represents what is shown in the animation.”

Screenshot:

- Click on the screen to view the animation of gas in a pressure chamber.
- You may watch the animation as many times as needed.
- Think about which principle of gasses the animation is describing.
- Select the correct principle of gasses from the choices provided:

  1. Law of Thermodynamics
  2. Avogadro’s Law
  3. Charles’ Law
  4. Boyle’s Law
10. Ratio of Inscribed Figures

Grade Range and Subject: Grades 9-12, Math

Target Features: Drag and drop, Images, Enabled elements

Description: Students are presented with a diagram of a triangle superimposed over a second one. The vertices of the second triangle are the midpoints of the sides of the first triangle. Students change the size and shape of the first triangle and observe the effect on the area of the second. They are then asked to explain the relationship between the areas of the two triangles as a result of changing their shapes.

Screenshot: