

Constructing Innovative Computer-Administered Tasks and Items According to Universal Design: Establishing Guidelines for Test Developers

Michael Harms¹

Kelly Burling²

Walter Way¹

Elizabeth Hanna¹

Bob Dolan²

Pearson Educational Measurement¹

Center for Applied Special Technology (CAST)²

Paper presented at the April 2006 Meeting of the National Council on Measurement in Education
(NCME)

cp0607

Abstract

Developing computer-based assessment items according to universal design principles requires understanding the test, as well as students' diverse special needs. The increased capabilities offered by new technologies offer new opportunities to deliver assistive technologies, but can also increase the range of factors that must be considered by test developers. These factors can interact in ways not easy to predict, with potential to affect test validity. This paper proposes the Universal Design for Computer-Based Testing (UD-CBT) framework for establishing guidelines that may be applied in designing non-standard computer delivered assessments consistent with the goals of Universal Design.

Introduction

The increasingly high-stakes associated with large-scale testing programs, especially those in the K-12 arena, have generated concomitant attention on how assessments are developed. Few would argue against the fundamental importance of the test development process in an assessment program. For example, The Standards for Educational and Psychological Testing (AERA, APA, NCME, 1999) includes 27 specific standards on test development and revision, and also state that issues of validity, reliability, and fairness are interwoven within the test development process. Most organizations in the business of developing tests have their own internal test development standards, and some make them available to the public (e.g., Educational Testing Service, 2002). Without question, thorough and clearly documented test development procedures are essential to the legal defensibility of a high-stakes testing program (Smisko, Twing & Denny, 2000).

Traditional test development procedures have been based on a well-established body of knowledge about how to write test items. These tend to be primarily aimed at multiple-choice items, but also address other item formats such as constructed-response and performance items (Haladyna, 1999; Osterlind, 1998). In recent years, however, requirements on large-scale assessments-- such as those associated with the No Child Left Behind Act of 2001-- have stressed the participation of populations with unique educational needs, varying cultural experiences, diverse linguistic backgrounds, and numerous special needs. To meet these

requirements, test development procedures have evolved to incorporate the concept of Universal Design (Johnston & Thurlow, 2002) and Universal Design for Learning (UDL; Rose & Meyers, 2001; Hall & Dolan, 2002; Thompson,).

UDL is predicated on individual learning differences, a concept supported by recent advances in brain research (see Rose & Meyer, 2002, for an overview). UDL asserts that learning occurs in and across three 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. Executive processes control the three networks by biasing or controlling how they contribute to behavior that is expressed in a particular situation. These networks and the pathways that connect them are different in all individuals. Variability in functional capabilities across and within these brain networks, and differences in executive processes are assumed to be a primary contributor to observable learning differences. However, learning differences do not necessarily result in predictable changes in academic performance. If instruction is structured in a way that aligns with the individual learning requirements then student performance can be maximized. UDL states that to the extent that instruction does not align with the learning needs of the student then a performance barrier is introduced and academic achievement may be reduced.

As extended to large-scale assessments, UDL principles uphold a test development process that facilitates participation of the widest possible range of students and results in valid inferences about performance for all students who participate in the assessment. The “universally designed test” should consider the needs of these students from the earliest stages of item development, and should involve choices in item development, test construction, and test administration that facilitate the most inclusive student participation possible, while still preserving the validity of the construct being measured. Thompson, Johnstone and Thurlow (2002) suggested seven elements that assessment professionals should consider when creating universally designed assessments. In combination the seven elements provide a framework for examining tests and their level of accessibility. Most large-scale assessment programs utilize many of these universal design elements to varying degrees. However, the goal of creating

inclusiveness sometimes competes with equally important goals of construct validity and score comparability in ways that are not always easy to sort out (Hanna, 2005).

In coming years, technology will change the relationship between universal design and large scale assessment development. Computer use for assessment and instruction continues to accelerate as technology becomes more powerful and readily available. Not only are computers becoming more common in usage but they also provide an opportunity to expand the manner in which we conduct our large-scale assessment systems. As assessments have transitioned to the computer, experts have begun to apply the concepts of Universal Design to consider how computer-based tests can be developed in ways that provide access for students with disabilities (Hansen & Mislevy, in press; Allan, Bulla, & Goodman, 2003; Thompson, Thurlow, & Moore, 2003; Thompson, Thurlow, Quenemoen, & Lehr, 2002). However, because computerized assessments have not yet progressed very far towards their potential, the application of universal design principles to constructing innovative online assessments is also at an early stage.

The purpose of this paper is to consider the application of universal design principles to computer-based assessments and to provide a framework for developing guidelines that may be applied in developing non-standard computer delivered assessments. Applying UD-CBT principles to computer based testing relies, in part, on the flexibility of digital media. Just as transferring a text book to a digital format does not immediately make it more accessible, transferring a traditional paper-and-pencil test to a computer does not necessarily make it more accessible. It is the flexibility of the digital format to incorporate multiple representations (text, video, audio), its ability to transform within and across media, and its ability to incorporate tools such as highlighters and linked dictionaries that makes it more accessible. By combining the potential of digital media and technology in CBT, a more accessible assessment can be created by incorporating multiple options for accessing and responding to assessment items.

The paper will be composed of three sections. In the first section, we introduce a theoretical basis for the UD-CBT guidelines framework through a discussion of construct validity. In the second section of the paper, we elucidate a UD-CBT framework to focus design recommendations based on three distinct categories of considerations: Student Characteristics, Interface Characteristics, and Information Content Characteristics. A framework for examining

Student Characteristics and Interface Characteristics is presented along with a discussion of how design recommendations are applied within this framework. In the final section, we describe features of the UD-CBT guidelines that can support the work of a test developer. These include the identification of barriers to student processing, the establishment of functional profiles of affected students, and the identification of CBT design solutions that can overcome these barriers. The relationship of the taxonomy to more general assessment considerations is discussed, along with areas for future research.

Framing UD-CBT in Terms of Construct Validity

In applying UD-CBT, parallels between the processes involved in learning and the processes involved in understanding and responding to test items become apparent. In both learning and assessment, the introduction of barriers can reduce student performance by creating construct-irrelevant challenges that alter the characteristics of the task. This problem is further complicated by the fact that these barriers can originate from a wide range of potential sources. Information structure, software interface characteristics, disabilities, and instructional practices are factors that can all interact to limit achievement if they conflict with student-specific needs. In computer-based assessment, these factors can also represent a significant source of task-irrelevant variability that can modify the intended test construct, resulting less valid interpretations of student performance. Recent technological advances in testing practices have created opportunities to deliver assessment items that measure learning in new and innovative ways. However, the added capabilities provided by computer-based testing technologies also have the potential to introduce unforeseen barriers due to interface characteristics and new types of media.

A central challenge for the design of CBT items is to identify design issues representing sources of task-irrelevant variability for student populations with different needs. At a practical level, a first goal of the UD-CBT framework is to create an easy-to-use reference for CBT Item developers to relate higher-level theoretical considerations to current and proposed test item designs. It should represent a classification system that organizes current research in a way that facilitates linking design guidelines to specific design contexts. Such a structure should also permit systematic evaluation of test item design proposals by identifying strengths and

weakness for specific student test populations. A second goal of UD-CBT is to extend UD foundations to include current human performance and interface design research relevant to computer-based assessment. A necessary requirement of the test item creation processes is the continual investigation of how students think about and respond to specific test item features. (Leighton & Gokiert, 2005) Although extensive research has been performed to examine the statistical properties of test items, less effort has been directed at understanding the potential for item features to introduce construct-irrelevant variance. (Haladyna & Downing, 2004; Ferrara et al, 2003; Leighton & Gokiert, 2005).

Construct irrelevant variance can be thought of as systematic measurement error resulting from unintended changes to an item construct due to the way it is implemented within a particular design framework or technology. Item level construct irrelevant variance results from reliably measuring one or more sources of variance that are irrelevant to the interpreted construct (Messick, 1994). This can consistently inflate or reduce scores for certain individuals or groups in ways that are unrelated to what the test is intended to measure. The overall validity of an assessment is intrinsically tied to both the construct validity of the individual items it contains, and the degree to which measures are valid across the test population. Therefore, minimizing construct-irrelevant variance at the item level should result in significant increases in the validity of a test.

Identifying and defining the sources of variance present in an item allows an assessment developer to consciously identify which are construct relevant and which are not. Some construct-irrelevant sources of variance can be reduced or eliminated by implementing design solutions. For those that cannot, developers may choose to construct alternate versions of the item for the students impacted. It is not possible to remove all construct irrelevant sources of variance from an assessment, but the extent to which developers are aware of their impact increases their ability to make valid interpretations of student performance. The result is a more universally designed assessment.

Within UD-CBT, the mechanisms by which construct irrelevant variance can impact test item validity can be thought to fall into two general categories:

Direct Threats to Validity – whether the right content is conveyed within a particular technology or medium (*expression of a construct*)

Indirect Threats to Validity – whether the interface interferes with the construct by adding additional skill or knowledge requirements due to the interaction between the student and the medium (*interface between student and construct*)

Direct threats to validity typically occur when the medium used to present a construct causes it to be altered in an unexpected way. For example, a question previously presented using only text-based description is modified to include both text and a multimedia component for a computer-based test. However, the information within the multimedia component that was chosen is somewhat ambiguous. As a result, the interpretation of the information conveyed by the multimedia component is impacted by the student's culture, previous learning, and previous experience. After viewing the multimedia component, the student believes the text and the multimedia are communicating different information. She decides that her original interpretation of the question from the text was wrong and answers the question incorrectly. In this case, the ambiguity within the multimedia component created a direct threat to validity by creating an unintended change in how the question was interpreted which caused an incorrect response.

In contrast, indirect threats to validity occur due to difficulties in *interacting with* the item content. For example, a particular computer-based test item requires that a student drag text and graphics to a response area and assemble a graphical representation of an ecosystem. The student has seen diagrams like this in class and has drawn them on paper, so she thinks she knows what her answer needs to look like. However, the test item requires that she put the parts of the diagram together in a different order than she's used to. She's also not sure what things she can move on the screen and where she can move them to. She spends two minutes learning how the interface works but starts to get concerned that she's going to run out of time. She hurries to finish her diagram but makes a mistake and incorrectly answers the question.

In this case, the expression of the construct was correct, but construct irrelevant variance was *indirectly* introduced due to the usage characteristics of the interface. The student was able to eventually learn how to use the interface, but her thinking and task flow were disrupted. As a result, she became anxious and made a mistake. Since she spent two minutes to learn the item interface, she would have had less time available to answer other questions on the test which

could cause other errors. In this case, the complexity of item interface represents a potential impact to validity because it takes time to learn, disrupts task flow, and makes learning the interface part of the construct that's measured.

Determining the exact point at which a test item design introduces unacceptable levels of construct irrelevant variance is not straightforward. The National Center on Educational Outcomes (NCEO) recommends that test item designers should “minimize knowledge and skills required beyond what is intended for measurement” (Thompson, et al, 2005). However, in the same discussion NCEO also notes that the level of skill required by a design “...depends on how discrete the standards are; minimal skills can be embedded in more complex contextualized items. Ultimately, it depends on what you are measuring.” (Thompson, et al, 2005) The appropriateness of an item design will in part be influenced by whether it is intended to measure isolated, basic skills or higher level thinking. In these cases, item validity and the determination of what constitutes a barrier will need to be considered in the context of the intent of the item.

Removing barriers to increase test validity requires a clear understanding of both direct and indirect threats to validity and their sources. Currently, test developers impact a limited number of these threats using practices such as test equating and scaling, differential item functioning analyses, bias reviews, and test accommodations, and they have begun to address some threats to validity at the item level through techniques such as the use of simplified language. As tests move to computers and include new types of media and multi-media the potential increases for individual items to present barriers not detectable through traditional reviews and analyses, particularly for low-incidence populations. In-depth item analysis will be even more vital to ensure that the intended constructs are being measured and that irrelevant sources of variance are minimized. This will not, however, ensure that all items will include only media that is accessible to every individual. A benefit of CBT is the opportunity to build multiple, flexible supports into tests at the item level. Video can include closed captioning, or rich description; text-to-speech and screen readers can provide auditory access to item content, and multiple language dictionaries can be made available for vocabulary support. Optimizing item design and deciding which supports are necessary and valid requires concrete understanding of the item's intent and the sources of variance each item component introduces. CBT can also

increase the validity of assessments for all students by providing an opportunity to individualize presentation, problem-solving, and response options.

Defining the UD-CBT Framework

To guide our efforts, we have established an organizing framework that separates factors impacting construct validity into three distinct categories: Student Characteristics, Interface Characteristics, and Information Content Characteristics. A brief description of these three categories of factors is included below.

Student Characteristics: Categories of Functional Capabilities, Disability Categories, and Language

Interface Characteristics: Interaction Model, Ease-of-Use, Color, Legibility, Navigation, and Status Information

Information Content Characteristics: Instructions, Test Item Content, Graphics, Multimedia, Language and Wording

The motivation for organizing the UD-CBT framework in terms of these factors is to allow solutions to be targeted at specific sources of problems. In addition to identifying threats to validity from student, interface, and information content characteristics, the UD-CBT framework also considers barriers resulting from the process that a student uses to interact with the test item. Specifically, our UD-CBT framework includes a Phases of Interaction Model that provides a theoretical basis for understanding how processing activities are sequenced, and their relationship to cognitive capacities. In this section of the paper, the definitions of student, interface, and information content characteristics will be discussed, along with the process of interacting with a test item.

Student Characteristics. Student characteristics include a range of human abilities to be considered when evaluating a test item design. The student characteristics defined in this section include an overview of primary functional limitations contributing to barriers that might impact assessment, and examples of tools and technologies that are commonly utilized to ensure that students with special requirements can participate in the general curriculum. To organize the

student traits that must be considered, different categories of possible cognitive, physical, and sensory limitations were defined that are representative of construct relevant and irrelevant factors that influence test item design. Table 1 describes this organizational structure and is comprised of various disability categories, the common functional limitations or barriers that students with these disabilities face, and accommodation technologies that can be used to overcome these barriers. It should be noted that, despite the structure of Table 1, we do not intend to place emphasis only on students with special needs. A central premise of both UDL and UD-CBT is that design solutions should be effective for the entire range of student needs. A design that is effective for a student with a specific need or functional limitation should also be effective for students without any special requirements. The category structure is intended to identify more detailed sources of variability associated with commonly defined special-needs populations to ensure 1) all needs of a particular population are being considered, and 2) in cases where gaps exist the need for a specific accommodation is noted.

To fully understand the construct-relevant and irrelevant aspects of individual items and the ways in which students interact with them requires a more detailed definition of the ways in which student differences are manifest while responding to assessment items.

Table 1: Description of Disability Categories

Disability Category	Common Functional Limitations	Accommodations Technologies
Blind	<ul style="list-style-type: none"> ▪ No Functional Vision, visual acuity 20/200 or greater 	Braille Embosser Refreshable Braille Display Nemeth Code Screen Readers/Talking Browsers Text-to-speech systems Optical Character Recognition Haptic Devices
Low Vision	<ul style="list-style-type: none"> ▪ Limited functional vision, visual acuity between 20/40 and 20/200 after correction 	Screen Magnification Screen Readers/Talking Browsers Text-to-speech systems Optical Character Recognition
Deaf / Hard of Hearing	<ul style="list-style-type: none"> ▪ No functional hearing, limited functional hearing ▪ Often corresponding delays in linguistic, social, emotional and cognitive development ▪ Literacy problems, especially delays in reading and writing, and difficulty with decoding and comprehension 	Volume Controls Signing Avatars Grammatical Support Tools
Physical Disability	<ul style="list-style-type: none"> ▪ Physical Mobility: fine motor skills, difficulty manipulating materials, limited eye movement ▪ Difficulty maintaining body positions, fatigue ▪ Potential for corresponding developmental brain disturbance 	Alternative Keyboards Alternative Mouse Systems Voice Recognition Systems Screen Readers/Talking Browsers Text-to-speech systems On-screen keyboards Word-prediction
Dysgraphia/Dyspraxia (fine motor issues)	<ul style="list-style-type: none"> ▪ Handwriting and drawing ▪ Writing fluency ▪ Manipulating materials ▪ Fine motor skills 	Alternative Keyboards Alternative Mouse Systems Snap-to constraints
Learning Disability: Reading/Language	<ul style="list-style-type: none"> ▪ Decoding, fluency, comprehension challenges ▪ Comprehending syntactic and semantic meaning ▪ Integrating information, making 	Grammatical Support Tools Text-to-speech systems

	<p>inferences</p> <ul style="list-style-type: none"> ▪ Connecting text ▪ Poor meta-cognitive skills ▪ Difficulty generating mental models needed for comprehension (reading, listening) ▪ Difficulty with written expression (planning, revising, self-regulating, writing mechanics) 	
Learning Disability: Math	<ul style="list-style-type: none"> ▪ Automaticity, fact retrieval ▪ Problem solving is interrupted due to concentration on fact retrieval ▪ Representations of word problems 	<p>Grammatical Support Tools Text-to-speech (MathML) Calculator</p>
Autism Spectrum Disorder: Asperger's Syndrome	<ul style="list-style-type: none"> ▪ Communication ▪ Listening comprehension ▪ Inferential reading ▪ Concept formation ▪ Problem solving ▪ Comprehending abstract concepts and language ▪ Ascertaining relevance ▪ Problem sensitivity ▪ Task switching ▪ Motor involvement ▪ Distractability ▪ Time perception ▪ Hypersensitivity (routines, light, sound, touch) 	<p>Grammatical Support Tools Text-to-speech systems</p>
Attention Deficit / Hyperactivity Disorder	<ul style="list-style-type: none"> ▪ Focus ▪ Sustaining attention ▪ Organization ▪ Task switching ▪ Time perception 	<p>Grammatical Support Tools Text-to-speech systems</p>
Mild Mental Retardation	<ul style="list-style-type: none"> ▪ Impaired functioning across subject areas ▪ Longer time to accomplish tasks, achieve mastery ▪ Generalizing skills ▪ Comprehension ▪ Expression ▪ Task Switching ▪ Time perception 	<p>Grammatical Support Tools Text-to-speech systems</p>

Interface Characteristics. An assessment item is often considered as a single 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 possible parts of an item and the sources of variance for each part. Then, it does not matter how the parts, or components, are combined to create new items. If the components of an item design can be identified, and each has sources of variance specific to it, an item analysis only requires determining which of the sources of variance are construct-irrelevant so the corresponding design recommendations can be implemented. Our initial efforts identified 11 test item interface components with which students interact while responding to CBT assessment items. The 11 components in our initial research are listed in Table 2.

Table 2: CBT Item Components

CBT Item Components	Definition
Text (Item Content)	Construct relevant terms or concepts in the instructions, stem, or stimulus materials.
Images	Photos, static images (artwork, maps, cartoons, etc.), icons (images on interface elements used to represent functionality), symbols (images that are commonly understood to represent a particular concept).
Audio	Independent audio recordings or an audio track accompanying a video or animation
Tables and Graphs	Tables used to organize information, convey structure and relationships. Graphs used to represent data visually
Mathematical Numbers and Symbols	Mathematical notation
Video and Animation	Visual representations that contain action
Response Options	Actions ranging from typing numbers or characters, clicking a box, clicking on a graphic or text, dragging icons or text while responding to item formats including selected response (multiple choice, multiple response, figural response [select part of a figure or graphic], ordered response [order or sequence a list of items in accordance with some rule]), sorting or categorizing problems or ranking items by correctness; constructed responses include typing a numerical answer to a quantitative question and figural responses where the student marks on, assembles, or interacts with a figure (build a circuit, plot points on a grid, correct errors in a passage).

Active Objects/Links	Words or icons that result in an action or take the user to a different location, Pictures with multiple active regions each which take the user to a different location
Multi-stage/Multi-part items	Multiple actions or responses required within one item. Screen elements or environment changes at each stage of multi-stage items, multi-part items have a different page for each part.
Constructed Response: Text	Language based composition ranging from fill in the blank to essays
Constructed Response: Math (Show your work):	Input a response ranging from a single number to complex proofs or displays of work.

Information Content Characteristics. The goal of UD-CBT is to identify design considerations that can be related to the construct relevant and irrelevant factors that impact test item validity. However, in many ways the construct definition is the least tangible aspect of an item design.

Table 3: UD-CBT Content Considerations

Content Consideration	Definition	Related Questions
Relevant	The item measures the content it intends to measure without extraneous content	<ul style="list-style-type: none"> ▪ Does the item clearly address knowledge, skills, or abilities identified in the test specifications? ▪ Is the content of the item clearly related to the objectives the item is supposed to measure?
Representative	Item content is aligned with test specifications Item elements correspond to materials and/or environments used in the classroom	<ul style="list-style-type: none"> ▪ Do the content and structure of the item align with how teachers and experts consider quality instructional methods? ▪ Does the item look like something students will have seen or used in the classroom?
Realistic	Unambiguous relationship between media or virtual environment and its real-world counterpart	<ul style="list-style-type: none"> ▪ If media is used to represent an actual process or event is it sufficiently realistic to be easily identifiable? ▪ Could it be mistaken for something else, or be too removed from an actual representation that matching the media to what it is intended to represent introduces construct-irrelevant

		cognitive load?
Synergistic	Item elements complement one another in conveying meaning	<ul style="list-style-type: none"> ▪ Do multiple elements stimulate the same processing category simultaneously, do they compete? ▪ Are there multiple simultaneous visual or auditory stimuli?
Clear and Unambiguous	Item intent and process to achieve it are conveyed clearly and with contextualization	<ul style="list-style-type: none"> ▪ Do the instructions clearly convey the scope and intent of the item? ▪ Are the steps necessary to reach the end, and how to proceed through them clear? ▪ Is the context sufficiently defined?
Free of Bias	Item is sensitive to the full population of test takers.	<ul style="list-style-type: none"> ▪ Is the item sensitive to cultural, socio-economic, gender, age, language, disability and regional issues? ▪ Will prior knowledge unfairly advantage one group over another? ▪ If post field-testing, does the DIF analysis indicate bias?
Consistent across items	Item format, tools, and operability are the same throughout a test. Students do not need to <u>learn</u> how to do something to respond to an item.	<ul style="list-style-type: none"> ▪ Are all elements presented in a manner consistent with other items (e.g., items appearing underneath text, test navigation along the bottom of the screen, blue underlined text indicates a link)? ▪ Are all tools located in the same space in the design? ▪ Are all tools and functionalities accessed in the same way as in other items throughout the test?
Appropriate Time and Task Load	The time required to view and interact with item elements has been considered and is appropriate to the intended difficulty and level of inquiry of the item. The impact of the item on student's time or energy to complete the rest of the test has been considered.	<ul style="list-style-type: none"> ▪ Is the length of any multi-media elements appropriate for the difficulty of the item and the level of inquiry (e.g. recall versus problem solving)? ▪ Is the time required to interact with an item appropriate for the difficulty and level of inquiry of the item? ▪ Do design features, such as multiple screens, increase task load inappropriately (e.g. multiple screens taxing working memory when the measurement focus is drawing inferences between text)? ▪ Is the time or task load of an item going to negatively impact the amount of time or energy a student will have to spend on other items?

The construct an item is designed to measure is conveyed through the item’s information content. The relevance of the information to the item construct, and the clarity with which it is presented directly impact item validity and the validity of the assessment. If the information is misaligned to the construct or is not clearly contextualized and presented, the interface design and student characteristics hardly matter because the item is invalid even before an interface is designed around it or a student interacts with it. Thus, high-level content considerations, such as those identified in Table 3, must play an important part in the UD-CBT framework.

UD-CBT Phases of Item Interaction Model. The UD-CBT framework addresses the issues associated with interacting with a test item from both the “what” and the “how” perspective. *What* students are presented and *how* they interact with what is presented both impact the usability and validity of an item. An examination of how a process is performed often reveals undiscovered issues that result from the interaction of individual factors in a dynamic context.

UD-CBT uses a traditional concept of stimulus-response interaction as a basis for examining CBT items from the process perspective. A model was identified that defines the information processing steps that are necessary to complete an assessment item. It is based on a logical progression of a student working through a test item from its initial presentation until completion of a required response. The 5-Stage UD-CBT Interaction Model is an expansion of this logical progression and incorporates cognitive, perceptual, and executive processing functions.

Within UD-CBT, interacting with a test item is thought to occur in three general phases:

- 1) **Item Presentation** – where the task of the student is to recognize and understand the information presented in the item prompt;
- 2) **Strategic Interaction** – where the task of the student is to manipulate, reorganize, modify or combine the information in the prompt with prior skills and strategies they have learned; and
- 3) **Response Action** – where the task of the student is to plan, organize, and construct a response of some kind to express what they know.

Within these three phases, there are steps that occur in which information is filtered, transformed, and integrated with existing knowledge. The processing that is performed at each step can be in one or more of six categories: perceptual, linguistic, cognitive, motoric, executive, and affective (See Figure 1). These six categories are not intended to be strictly representative of

biologically distinct information processing systems, nor are they free of overlap. They are meta-constructs that define the range of conceptually distinct processing that can occur throughout the process of interacting with an item.

The six categories of processing frame the questions that must be applied to items in order to understand the ways in which items function differently for different students. They broaden the question, “how do items differentiate between students,” by suggesting that we need, instead, to ask how items discriminate according to the perceptual, linguistic, cognitive, motoric, executive and affective processing they require.

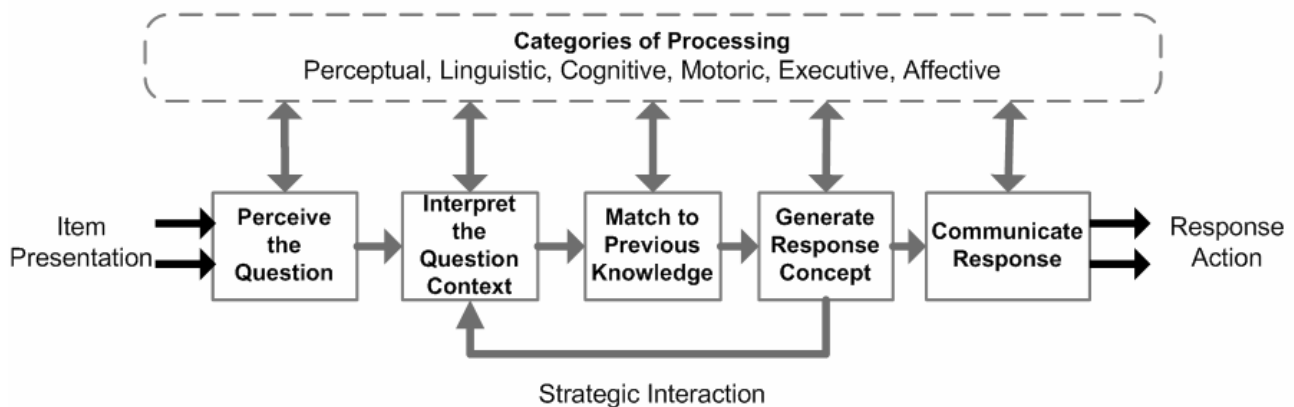


Figure 1. UD-CBT Phases of Item Interaction Model

UD-CBT assumes a series of stages but acknowledges that a wide range of processing occurs in parallel throughout the phases of item interaction. It assumes that information is processed simultaneously by several different parts of the memory system (Rumelhart and McClelland, 1986) and that the results of distributed processing are combined into higher-level information representations.

UD-CBT also incorporates four general principles shared among most modern theories of human information processing (Huitt, 2003). The first is the *assumption of a limited processing capacity* within sensory and cognitive systems. This means that the amount of information that can be processed by the system is constrained in observable and predictable ways. Bottlenecks, or restrictions in the transmission and processing of information, frequently occur at very specific points and under specific conditions.

A second principle is that a *control mechanism is required* to oversee the encoding, transformation, processing, storage, retrieval and utilization of information. That is, not all of the processing capacity of the system is available; an executive function that oversees this process will use up some of this capability. When one is learning a new task or is confronted with a new environment, the executive function requires more processing power than when one is doing a routine task or is in a familiar environment.

A third principle is that there is a *two-way flow of information* as we interpret information about of the world around us. We use a combination of information captured through the senses (often referred to as bottom-up processing) and information we have stored in memory (often called top-down processing) in a dynamic process as we construct meaning about our environment and our relations to it. As a result, our knowledge and experience shape our perceptions of our current environment.

A fourth principle generally accepted by cognitive psychologists is that the human organism has been *genetically predisposed to process and organize information in specific ways*. For example, it is widely accepted that processing associated with language is functionally distinct from processing that occurs on pictures and imagery. This is supported by the fact that deficits in one skill do not necessarily cause corresponding reductions in all others. Therefore, categories can be created that represent relatively independent forms of processing that occur during the performance of complex tasks.

In addition, a fifth principle assumed by UD-CBT is that there is *individual variability* in the level of performance across different categories of information processing. The six categories of processing defined within UD-CBT contribute to test performance in different ways and at different levels depending upon individual student characteristics. This variability is assumed to result from both the biological predispositions inherent to the individual and the effects of situation specific influences.

In combination, the theoretical considerations embodied by the phases of interaction model and the information processing principles associated with it form a structured basis for organizing research and test item design options. Existing research and guidelines for software interface design can be assessed within the UD-CBT theoretical framework to identify the ways that they relate or do not relate to computer-based testing. As a result, relevant research can be integrated into the

existing body of knowledge about computer-based assessment to accelerate the evolution of test item designs.

Applying UD-CBT to Test Item Design. The categories that have been created suggest the questions to be asked about each component of a test item design. It is important to define, as specifically as possible, the individual factors within each category that affect student interaction with each component. For instance, “How does a student interact with an image on a perceptual level? What distinguishes between students who perceive an image well and those who do not?” The individual factors constitute sources of variance. Sources of variance define the ways in which students differ in interacting with and responding to item content, the media through which it is conveyed, and the physical interface.

From a process perspective, the presentation of the item and the response options are where the interface issues are most distinct. However, the higher-level information processing activities that occur throughout the item interaction process all have potential to introduce variance. Factors such as cognitive load, anxiety, and differences in media-specific information processing requirements interact in ways that are not directly observable. Time requirements and the efficiency with which a student can interact with an item can also have subtle impacts on validity at both an item and overall test level.

In combination, the UD-CBT categories along with the Phases of Item Interaction form a basis for evaluating the strengths and limitations of a design guideline. The categories provide a definition of the more concrete aspects of interacting with a test item (“what”) and the Phases of Interaction Model provides a method for thinking about how these factors interact throughout the dynamic process of completing a test item (“how”)

Understanding the variety of issues to be considered requires a flexible, systematic approach that can evaluate a design from multiple perspectives and at different levels of detail. Designers often use combinations of checklists, rating scales, guidelines, and structured review processes to direct and support design efforts. However, the effectiveness of the process can be enhanced or limited by the quality and appropriateness of the approach that is used. The category structure of UD-CBT ensures that guidelines are appropriate to the problem context and framed at the appropriate level of detail.

Information Structure within UD-CBT. The information structure used by UD-CBT can be thought of as a 3-dimensional matrix within which information is organized. Categories defined for Item Components, Student Groups, and Categories of Processing represent the 3 dimensions used to locate design data. (see figure 2) Each of the three categories includes the elements specified in the UD-CBT theoretical model and also includes an additional summary category (i.e., All Item Components, All Student Groups, and All Categories of Processing). The summary category represents higher-level recommendations that can be applied across all subcategories.

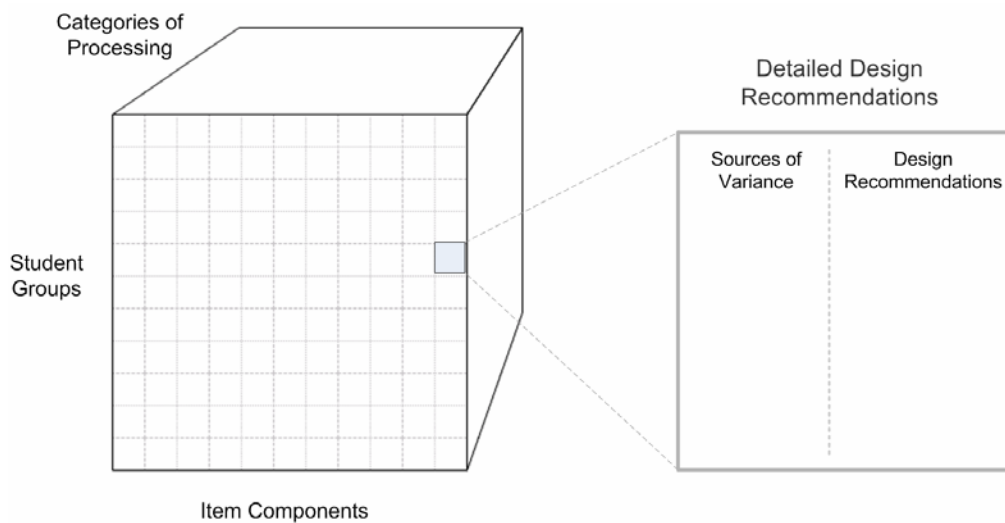


Figure 2. UD-CBT Information Structure

Within each cell of the matrix, the sources of variance associated with that particular combination are noted, along with design options or guidelines. At the current stage of UD-CBT's development, the sources of variance within each cell would identify *indirect threats to validity* resulting from CIV related to student and interface component factors. In cases where research identified no sources of variance or no available research exists, it would be noted in the appropriate cell. *Direct threats to validity* relating to the expression of a construct would be assessed using a combination of the UD-CBT Content Checklist and traditional content review processes.

Evaluating Test Item Designs within UD-CBT. Analyzing a test item design within UD-CBT is fairly straightforward and could consist of the following steps.

- 1) **Evaluate the item design for *Direct Threats to Validity*** - This analysis determines whether the right content was conveyed within the particular technology or medium as it is defined in the test item specification. It consists of the following two steps:
 - a. Assess the construct using the UD-CBT Content checklist
 - b. Assess the construct using traditional item review processes
- 2) **Evaluate the item design for *Indirect Threats to Validity*** – This analysis determines whether the interface interferes with the construct by adding additional skill or knowledge requirements due to the interaction between the student and the medium.
 - a. Identify the *Level of Analysis* to be performed – Choose guidelines for “All Students” to find Universal Design Guidelines applying to the entire student population. For focused design guidelines choose a specific student population of interest.
 - b. Identify *Item Components* within the test item design - Choose the specific item component to be evaluated or choose “All Item Components” to find general design guidelines.
 - c. Evaluate the item design against *Detailed Design Recommendations* for each *Source of Variance* within each *Category of Processing* – Evaluate the item design using the Sources of Variance listed for each of the six Categories of Processing.
- 3) **Revise the item design to incorporate relevant design recommendations and re-evaluate using the UD-CBT process.** - Upon completion of design changes identified through UD-CBT the same process should again be applied to ensure that no additional Sources of Variance were introduced.

The outcome of the UD-CBT analysis process would be the identification of design elements of the test item design likely to represent sources of variance and proposed methods for minimizing that variance. While no design evaluation can definitively guarantee that all potential sources of construct-irrelevant variance are eliminated, it is expected that a systematic, comprehensive evaluation based on research and best-practices will identify a greater number of issues early. This will facilitate the creation of computer-based assessment items that are accessible to the widest range of students and more completely achieving the goals of Universal Design.

Conclusions

Effective design of innovative computer-based test items leveraging new technologies requires consideration of issues from multiple perspectives. The increased opportunity to create items effective for a wider range of test populations also increases the number of direct and indirect threats to validity that must be considered to insure that construct-irrelevant variance is not

inadvertently introduced. The addition of new capabilities and the availability of new media types requires guidelines and item review methodologies able to identify sources of construct-irrelevant variance not present in traditional paper based assessment.

The UD-CBT framework was proposed as an organizing structure to create and apply research-based guidelines to assessment design. The conceptual distinction between student factors, interface factors, and information content factors was proposed as a basis for considering categories of sources of construct-irrelevant variance. The UD-CBT Phases of Interactivity Model was presented to provide a theoretical basis for assessing and understanding the process of interacting with a test item. The six categories of processing listed within the phases of interaction model (i.e., perceptual, linguistic, cognitive, motoric, executive, and affective) were identified as a way to organize sources of variance for each item component. As a whole, UD-CBT represents a comprehensive framework for analyzing and understanding a wide range of factors that influence construct validity.

References

Allan, J. M., Bulla, N. and Goodman, S.A. (2003). *Test Access: Guidelines for Computer-Administered Testing*. American Printing House for the Blind: Louisville, KY. Available from: <http://www.aph.org>.

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

Denckla, M. B. (1996). A theory and model of executive function: A neuropsychological perspective. In G. R. Lyon & N. A. Krasnegor (Eds.), *Attention, memory, and executive function* (pp. 263–278). Baltimore, MD: Paul Brookes.

Dolan, R. P. and Hall, T. E. (2001). Universal design for learning: Implications for large-scale assessment. *IDA Perspectives*, 27(4), 22–25.

Educational Testing Service (2002). *ETS standards for quality and fairness*. Princeton, NJ: Educational Testing Service.

Ferrara, S., Duncan, T. G., Perie, M., Freed, R., McGivern, J., & Chilukuri, R., (April, 2003) *Item Construct Validity: Early results from a study of the relationship between intended and actual cognitive demands in a middle school science assessment*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Haladyna, T. (1999). *Developing and validating multiple choice items*. Hillsdale, NJ: Erlbaum.

Haladyna, T. M., & Downing, S. M., (2004) *Construct-irrelevant variance in high-stakes testing*. *Educational Measurement: Issues and Practice*, 17-27.

Hanna, E. I. (2005). *Inclusive Design for Maximum Accessibility: A Practical Approach to Universal Design* (PEM Research Report 05-04). Iowa City, IA: Pearson Educational Measurement.

Hansen, E. G., & Mislevy, R. J. (in press). Accessibility of computer-based testing for individuals with disabilities and English language learners within a validity framework. In M. Hricko & S. Howell (Eds.), *Online assessment and measurement: Foundation, challenges, and issues*. Hershey, PA: Idea Group Publishing, Inc.

Huitt, W. (2003). The information processing approach to cognition. *Educational Psychology Interactive*. Valdosta, GA: Valdosta State University. Retrieved March 19, 2006 from, <http://chiron.valdosta.edu/whuitt/col/cogsys/infoproc.html>.

Leighton, J.P., & Gokiert, R. J. (2005). *The Cognitive Effects of Test Item Features: Informing Item Generation by Identifying Construct Irrelevant Variance*. Paper presented at the Annual Meeting of the National Council on Measurements in Education (NCME), Montreal, Quebec, Canada (April 2005).

Osterlind, S. J. (1998) *Constructing Test Items: Multiple-Choice, Constructed-Response, Performance and Other Formats*. Norwell, MA: Kluwer Academic Publishers.

Rose, D. and Meyer, A. (2000). Universal design for learning, associate editor column. *Journal of Special Education Technology*. 15 (1). Available from: <http://jset.unlv.edu/15.1/asseds/rose.html>

Rumelhart, D., & McClelland, J. (Eds.). (1986). *Parallel distributed processing: Explorations in the microstructure of cognition*. Cambridge, MA: MIT Press.

Smisko, A., & Twing, J. S. (2000). The Texas model for content and curricular validity. *Applied Measurement in Education*, 13(4), 333-342 .

Thompson, S., Thurlow, M., & Moore, M. (2003). Using computer-based tests with students with disabilities (Policy Directions No. 15). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved July 20, 2005, from: <http://education.umn.edu/NCEO/OnlinePubs/Policy15.htm>.

Thompson, S. J., Thurlow, M. L., Quenemoen, R. F., & Lehr, C. A. (2002). Access to computer-based testing for students with disabilities (Synthesis Report 45). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved July 20, 2005, from: <http://education.umn.edu/NCEO/OnlinePubs/Synthesis45.html>.

Thompson, S. J., Johnstone, C. J., and Thurlow, M. L. (2002). Universal design applied to large scale assessments (Synthesis Report 44). Minneapolis, Minn.: University of Minnesota, National Center on Educational Outcomes. Retrieved January 9, 2004, from: <http://education.umn.edu/NCEO/OnlinePubs/Synthesis44.html>.