Graduate Management Admission Council®

Designing Templates Based on a Taxonomy of Innovative Items

Cynthia G. Parshall
Measurement Consultant
and
J. Christine Harmes
James Madison University

Keynote Address Presented on June 7, 2007

2007 GMAC® Conference on Computerized Adaptive Testing

Abstract

The design of high quality innovative item types is a challenging but important measurement task. This paper addresses the use of a taxonomy and item templates as related elements that test developers can consider when designing and developing innovative item types. The taxonomy for innovative types includes the seven dimensions of: assessment structure, complexity, fidelity, interactivity, media inclusion, response action, and scoring model. Item templates are a structured means of collecting and storing item information that can be used to improve the efficiency and security of the innovative item writing process. Specific item types can be evaluated in terms of the impact of their taxonomic levels on test development issues. such as exam program construct, psychometrics, programming needs, examinees' computer skills, and costs. Finally, once innovative item types have been selected and designed, item templates can also be developed to guide and support the item writing process.

Acknowledgment

Presentation of this paper at the 2007 Conference on Computerized Adaptive Testing was supported in part with funds from GMAC®.

Citation

Parshall. C. G. & Harmes, J. C. (2007). Designing templates based on a taxonomy of innovative items. (2007). In D. J. Weiss (Ed.). ***Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing.*** **Retrieved [date] from** www.psych.umn.edu/psylabs/CATCentral/

Author Contact

Cynthia G. Parshall, 415 Dunedin Ave., Tampa, FL 33617, cparshall@tampabay.rr.com or J. **Christine Harmes, 821 S. Main St., MSC 6806, Harrisonburg, VA 22807**, harmesjc@jmu.edu

Designing Templates Based on a Taxonomy of Innovative Items

Innovative assessments have great potential for expanding the content areas and construct coverage of an assessment. However, there are undoubtedly challenges and risks in adding innovative items to an exam program, and these must be considered by test developers who are considering this step. Furthermore, the selection of types of innovative items should be carefully made, with a focus on identifying those innovations that will be of greatest benefit to the specific program. Once these decisions have been made, procedures and guidelines for the actual item writing process will be of critical value. Item templates are tools that can streamline the item design and development process, and thus increase efficiency.

This paper provides a set of materials to help test developers who are considering the use of innovative item types. First, we provide an updated, revised taxonomy for innovative assessments. This taxonomy will provide detail regarding the various design decisions and trade-offs that test developers must address when developing innovative items. Before considering the use and specification of item templates, dimensions of the taxonomy should be carefully considered. Following a description of the taxonomy, along with relevant examples, we discuss the design of item templates. Within the discussion of item templates, we provide a brief section addressing the test development issues of: exam program construct, psychometrics, programming needs, examinees' computer skills, and costs.

A Taxonomy for Innovative Assessments

Several ways of categorizing innovative items have been proposed (Koch, 1993; Luecht & Clauser, 2002; Parshall, Stewart, & Ritter, 1996; Zenisky & Sireci, 2002). One taxonomy for innovative items (Parshall, Davey, & Pashley, 2000) provided a comprehensive framework for innovative item types in terms of five dimensions: item format, response action, media inclusion, level of interactivity, and scoring method, or algorithm.

With the rapid advancement of technology has come increasingly complex options for including innovations in testing. In particular, as innovative assessments have developed it has become increasingly inaccurate to refer to all of them as "items." These changes necessitate a refinement to the taxonomy of testing innovations, in order to allow for assessment applications that are increasingly divergent from the traditional testing environment. The revised taxonomy provided in this paper primarily expands the dimension of "item format" into the broader term "assessment structure," while also adding the new dimensions of complexity and fidelity. The seven levels of the revised taxonomy are: assessment structure, complexity, fidelity, interactivity, response action, media inclusion, and scoring algorithm.

Assessment Structure

The term assessment structure is used to encompass a broader range of what has traditionally been called the item format. Computer-based assessments have the potential to extend beyond discrete items, through situated tasks, to simulated environments (Harmes & Parshall, 2005). The majority of both traditional and innovative items that have been used operationally could be classified as discrete items. Further along the assessment structure continuum is the situated

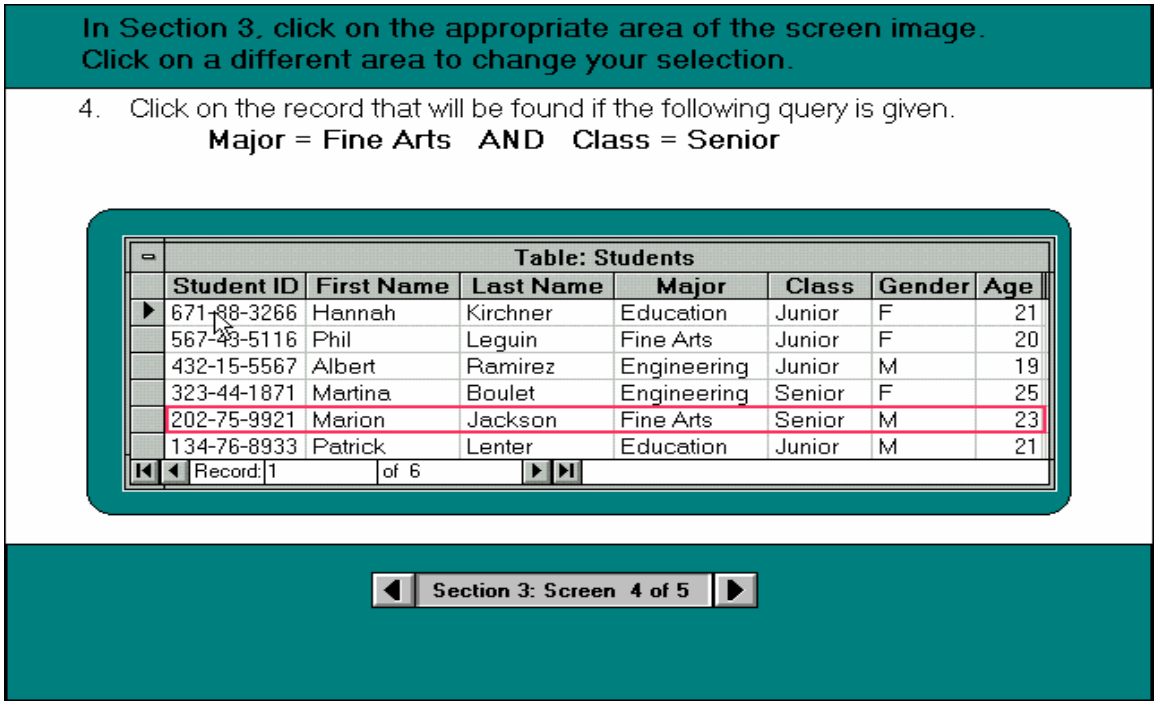
task. A situated task consists of the presentation of a series of actions related to a single scenario. At the far end of the assessment structure continuum is the simulated environment. A simulated environment is a fully elaborated, computerized presentation or mediation.

Discrete items. Most computerized assessments are comprised of discrete items. The primary two categories of discrete items are selected response items and constructed response items. A variety of innovative item types have been developed within both of these categories. The true value of these innovative item formats is their potential to improve measurement. Several formats are intended to reduce the effect of guessing by expanding the range of possible responses. Other formats expand the content or cognitive areas that can be measured by a test.

In one simple computerized adaptation of the selected response multiple-choice item, examinees might be asked to click on and select the proper sentence from a reading passage or to select one part of a graphic image. This type of innovative item can potentially improve measurement both by reducing guessing and by affording more direct measurement. Examinees interact directly with the entire reading passage or image, rather than to a lettered, indirect subset.

The "hot spot" or figural response item is an additional extension of the selected response item type. In figural response items, examinees respond by selecting a part of a figure or graphic. For example, an examinee might be asked to select a specific element or area within a spreadsheet, or a location on a chest diagram. An example of a figural selected response item can be seen in Figure 1.

Figure 1. A Sample Figural Selected Response Item



Another selected response item type frequently adapted to computer-based tests (CBTs) is the multiple-response type. In this item type, an examinee is asked to select more than one option; the examinee might be asked to select a specified number of options, or to select "all that apply." A further instance of selected response innovative item types is the ordered response type. In this instance, examinees are presented with a list of elements that they are then asked to place in order or sequence.

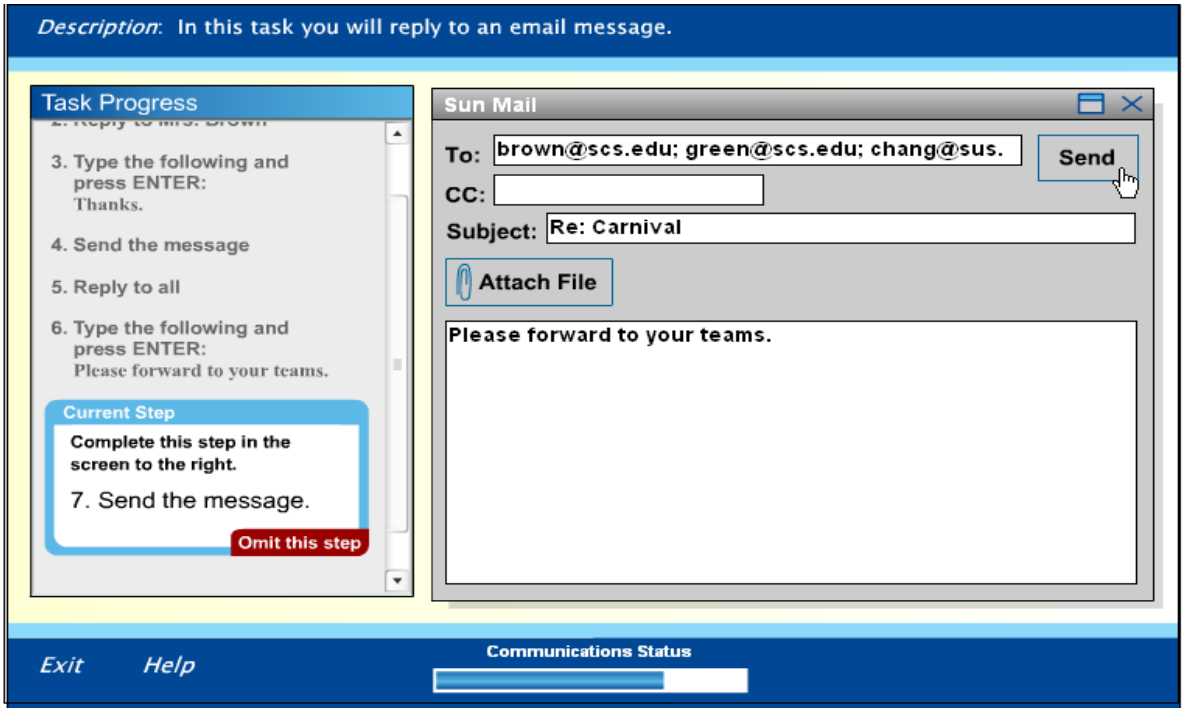
A wide range of constructed response items has been considered, varying from fairly simple formats that are easy to score, to far more complex formats that require the use of elaborate scoring algorithms. The simplest examples of constructed response innovative items are those that require the examinee to type a numerical answer to a quantitative question (e.g., O'Neill & Folk, 1996), or a very short response to a verbal question. In a slightly more complicated item type, examinees mark on, assemble, or interact with a figure on the screen (e.g., Martinez, 1993; French & Godwin, 1996). As constructed response items move towards more elaborate, integrated assessment structures they might be classified as situated tasks or possibly even simulated environments.

Situated tasks. Situated tasks use technology to present a realistic situation or scenario in which examinees are asked to create a product or solve a problem, typically through a series of actions or steps (Harmes & Parshall, 2005). While in some cases similar information can be collected through the use of discrete multiple-choice items, situated tasks provide the cohesiveness of operating within an integrated context. A distinction is often made as to whether the situated task is designed to progress in either a structured or unstructured manner. These design options have also been referred to as *path* and *open simulations*, respectively (Heffner, Knapp, & Rosenthal, 2004). A structured (path) task will progress through the same steps in the same order for all examinees. An unstructured (open) task will allow examinees to largely determine their own paths, based on any number of choices that may be made.

An example of a situated task that was designed to have structured progression is the Teacher Technology Skills assessment (Harmes et al., 2004) **.** In what is referred to as the "simulation" portion of the assessment, examinees interact with simulated versions of various software applications. A software-related situated task is presented to the examinees, along with a series of ordered steps that must be completed, in a specified sequence, to satisfy the task. Examinees are given the option of omitting a particular step, if they so desire. While the omitted step is scored as incorrect, this nevertheless allows the examinees to continue progressing through the remaining steps and thus to complete the rest of the task (see Figure 2).

These situated tasks have a surprisingly high level of fidelity, based simply on the fact that the content area being assessed is computer-based software, and thus the CBT software can readily emulate the real-world environment. Although fidelity is not essential for good measurement, in this instance targeting greater fidelity appeared to be an appropriate way of targeting greater construct relevance.

Figure 2. A Sample Situated Task
(© Florida Department of Education. Reproduced with permission.)



An example of a situated task designed to use unstructured progression is the credentialing test created by the National Council of Architectural Registration Boards (NCARB), as described by Braun (1994). This test includes a performance or simulation component, during which examinees respond to several "vignettes." Each of the vignettes presents an architectural problem or assignment. The examinee must use computerized drawing tools to design a solution to the problem within specified criteria. These problems have clear similarities to the constructed figural response item type, although the tasks, and the procedures for scoring responses, are far more complicated.

Another example of unstructured progression can be found in the computerized certification exam for accountants. (AICPA, 2003; AICPA, 2004). This exam includes situated tasks that assess research, judgment, and analytical skills with greater authenticity and a higher level of similarity to the actual job situation than the former exam provided. These tasks present a problem, provide appropriate resources, and require the examinee to perform tasks such as completing a tax form or searching a database of accounting literature and then copying and pasting appropriate resource material to assemble an auditor's report.

"Patient case simulations" were added to the United States Medical Licensing Examination (USMLE) in 1999 as a way to measure the patient management skills of physicians. In each of these situated tasks, the examinee is presented with a brief description of a patient who is complaining of some malady. The examinee can order medical tests or procedures, interpret the results of those procedures, diagnose the condition, and monitor changes in status over time and

in response to actions taken. All of the steps taken by the examinee are recorded, along with the simulation-based time in which each action occurred. Examinee scores are based on elements such as efficiency, thoroughness, timeliness, and avoidance of unnecessary, risky, or dangerous actions. Analyses indicate that these simulated cases are able to measure something more (i.e., patient management skills) than was possible with traditional multiple-choice items (Dillon et al., 2004).

Simulated environments. The most elaborate level of assessment structure is that of simulated environments. A simulated environment provides a performance assessment, conducted through the use of technology, to simulate extensive components of a real-world environment. The key difference between situated tasks and simulated environments is that the simulated environment assessment is presented in a setting in which most of the elements of interacting in the real environment are replicated (Harmes & Parshall, 2005). This replication might involve testing in a real, or simulated, version of the physical environment (e.g., emergency room or military tank), and might also include real or simulated tools and personnel that would be part of the actual environment. Although there are few operational examples of simulated environments in current practice, this is likely to be an area of future development.

Tsai et al. (2003) described an assessment designed to assess pediatric residents' response in an emergency room situation through the use of a high-fidelity child mannequin. The mannequin includes correct anatomy and is capable of producing correct responses (such as heart, breathing sounds, pulse). The assessment is conducted within a simulated environment employing real medical monitoring equipment, real medications, and additional personnel acting as "staff" and patient's "family." The mannequin's technology has some limitations, including an inability to produce movement, skin temperature, or color changes; when these physical changes are intended to have occurred, examinees are orally informed.

Perhaps the most sophisticated and long-used simulations in assessment and selection are in the aviation industry. Flight simulators are used for selection, licensure, and training of pilots. The simulator's flight controls are duplicates of those in the actual cockpit of the airplane on which the examinee is being tested, and are calibrated to perform in a manner highly similar to the actual controls. A screen in front of the pilot simulates the view out of the cockpit window. Motion is simulated through the use of hydraulics. The examiner sits behind the pilot in the simulator and uses a control panel to direct the actions of the simulator, while reading from a script to simulate radio communications. An examiner causes the simulator to impose several emergency or abnormal situations. Although each emergency or abnormal situation is given one at a time, any incorrect actions by the pilot can cause additional corresponding difficulties to occur, thus illustrating the open nature of the simulation. For example, an engine fire upon take-off requires use of a standard emergency procedure. If the procedure is followed improperly, such as by shutting off the fuel to the normally operating engine instead of the engine that is on fire, the pilot then has to correct the new emergency that he or she just created in addition to the original emergency given by the examiner (Capt. Sean Reynolds, personal communication, May 9, 2005).

Advantages of the flight simulator include cost savings and safety. Even though the flight simulator is expensive technology, it represents a substantial operational cost savings over the

real environment. Furthermore, the simulator allows pilots to train and be tested on responses to situations that would be too dangerous to do in an actual airplane. Disadvantages tend to consist of relatively minor defects in terms of fidelity. These include infidelities in radio communications and in sensations such as motion, pressurization, and smell (Longridge et al, 2001; Capt. Sean Reynolds, personal communication May 9, 2005).

Simulated environments and situated tasks have far more elaborate assessment structures than discrete innovative items. These more elaborate assessments are also likely to need a great increase in the amount of time required to write and program the assessments, as well as an increase in the amount of examinee time required to complete these tasks. However, in many cases, simulated environments afford measurement of constructs or facets of constructs not possible with any other type of measurement tool. In highly critical assessment situations, such as life-threatening or other high risk situations, the time and expense involved might be of sufficient value to consider a simulated environment assessment.

In general, as assessments become more complex, they are likely to increase the test development effort. The more elaborate assessments will typically take longer to develop, be more expensive to program, and require a more extensive validation effort. Although the potential exists for increasing test validity by adding innovative assessments that address content or construct areas previously unmeasured, there are certain measurement risks as well. When an assessment concentrates on greater *depth* of measurement, this sometimes comes at the cost of reducing *breadth*. For example, when a test has fewer items overall, this can reduce construct representation and decrease estimates of validity. Furthermore, a reduced number of items can reduce test reliability. More innovative assessment structures might also create item writing challenges. Finally, the way in which the assessments are represented on the screen might also have implications. If the user interface is too complex it can require higher levels of computer skills than the examinee might have. A spuriously complex user interface can also contribute to construct irrelevant variance.

Complexity

For a computerized assessment, complexity can be defined as comprising the number and variety of elements that an examinee needs to consider when responding to an item. This includes both conceptual and functional aspects, as an item can include both onscreen elements that need to be interpreted as well as item components that an examinee might use. For example, a complex innovative item might include informative text or graphics in several different locations on the screen, as well as various functional elements such as active buttons, tabs, media players, or more.

Innovative items span a wide range of complexity. A low level of complexity is evident in a multiple-choice item with a simple, non-interactive graphic in the stem. This type of item would require little interpretation or inference beyond that required by a traditional item type. The difficulty of an item such as this is likely to be based almost entirely on the problem posed in the stem. Complexity is increased in an item type as additional visual elements are included on the screen. These might include text in headers, labels, tabs, or retrievable data. Other types of visual elements include graphics, whether static or dynamic, and icons. All of these forms of visual information need to be processed by the examinee, and thus affect the complexity of the

item. The examinee's task is also made more complex as active or functional elements are added. The inclusion of a single functional component, such as a media player, might only increase the complexity slightly, but when numerous active elements are included, the task can become substantially more complex. The highest levels of complexity are perhaps evident in extended response assessments, as these tend to include numerous, varied screen elements that an examinee might need to interpret or use.

Complexity might tend to increase as the interactivity or fidelity of an item increases. Furthermore, increased item complexity might be associated with most contextualized, integrated assessments. It is important to note that in many cases increased complexity is likely to be associated with an increase in the item's cognitive challenge. From a measurement perspective it is thus critical that the complexity of an item be construct-relevant. It will be important for test developers to carefully consider the complexity of an item type to ensure both that it is content-relevant and that it is targeted at an appropriate level. Furthermore, a spurious aspect of item complexity can also arise from an inappropriate complexity of the software's user interface. Just as we do not want traditional items to be made more difficult by "tricky" wording, so we should avoid artificial challenges due to poor usability of CBT items.

Fidelity

In the context of computer-based testing, fidelity can be defined as the degree to which the assessment provides a realistic and accurate reproduction of the actual objects, situations, tasks or environments that are part of the construct being measured (Harmes & Parshall, 2005).

The closer an item, task, or environment approximates the actual construct being measured, the higher the fidelity. Various levels of fidelity can be illustrated with examples related to the measurement of the skills of a commercial airline pilot for handling an emergency or abnormal situation. A very low fidelity example would be a text-based multiple-choice item that presents a situation and asks the examinee to select which of four actions should be carried out first. This is an appropriate level of fidelity to use when targeting an assessment of knowledge-level skills. However, it does not provide an indication of how well the pilot would perform in a real-world situation. A computer-based flight simulator, comprised of a software program on a personal computer and input devices such as a mouse or joystick, would add a further degree of realism, and would thus increase the fidelity. An example of an assessment with very high fidelity would be a performance test in a full flight simulator in which the simulated airplane environment has been replicated to appear and function as if it were a real airplane. This higher level of fidelity is appropriate in this application, despite the additional expense, because including a realistic context within the assessment of these professional skills is critical to public safety.

As the level of fidelity increases across this set of examples, so too does the investment of time and money required for development. Decisions about which level of fidelity to target will depend upon the purpose of the assessment. While a test in a computer-based flight simulator might provide enough information to differentiate between candidates for selection, a full-flight simulator with far greater fidelity would be necessary for qualifying a pilot to command an airplane. However, targeting a higher level of fidelity is not always recommended. Test developers should carefully match fidelity levels to desired score inferences.

Developing an innovative assessment with a high degree of fidelity clearly has certain challenges. One potential challenge to the validity of an assessment is the risk of targeting a high fidelity match to one environment, while providing a poor match to another environment. As with complexity and interactivity, assessments with greater fidelity are likely to be more expensive to program. In addition, a high fidelity assessment might require that specific computer hardware or monitors be available for test administration. It will be important for an exam program to target a useful level of fidelity, and not to divert program resources by exceeding that level.

Interactivity

Interactivity, as a facet of this taxonomy, describes the extent to which an item reacts or responds to examinee input. In this case, it does not refer to the adaptive nature of a computerized adaptive test (CAT). Interactivity can be specifically incorporated into some of the discrete innovative items, while the more elaborate assessment structures tend to use interactivity quite extensively.

The level of interactivity in an assessment often increases as assessment structure moves away from discrete, text-based, selected response. For example, most assessments at the discrete item end of the assessment structure continuum are likely to have minimal interactivity. For many of these basic innovative items, the only form of interactivity provided by the computer is a highlighted or shaded display of the response option selected by the examinee. At the next lowest level of interactivity, a few item types provide a limited type of informative feedback, increasing item-examinee interaction slightly. With these item types, the examinee acts and the computer then responds with some sort of reaction or information. For example, when examinees click on a histogram, scale, or dial they might see the bar or gauge move to reflect this selection (O'Neil & Folk, 1996). Examinees who edit text might see the new text embedded within the original passage, letting them re-read the passage with the change reflected (Breland, 1998; Davey, Godwin, & Mittelholz, 1997). Examinees who specify an order for a set of elements might see the elements rearranged in the new sequence (O'Neil & Folk, 1996). Finally, examinees who indicate the points on a grid where a line or curve should be plotted, might see that line or curve displayed (Bennett, Morley, & Quardt, 1998). Examinees can use this limited, informative level of interactivity to help them decide whether their initial selection had the desired effect. If it is important, within a given test content area, to incorporate this type of information into the assessment, then the inclusion of interactivity is appropriate.

More sophisticated use of interactivity typically occurs in assessment tasks that are situated within a representative context. Generally, as the situation increases in realism, more interactivity is likely to occur. One example of a situated task that includes a higher level of interactivity is a computerized research skills assessment (Harmes & Parshall, 2000). In this research skills assessment, examinees are asked to first formulate a search strategy related to a given research topic, and then to order the set of article titles that resulted from the literature search. As an additional level of interactivity provided in this task, examinees are able to click on each article title to see a complete citation and abstract, in a manner that is similar to a real literature search. The increase in interactivity allows for greater realism in the situated task, and thus provides an opportunity to gather information about the examinee's thought processes

within a realistic situation of conducting a literature search. The increase in interactivity provides an additional context that is appropriate for the construct being assessed.

In more elaborate innovative assessments, the entire setting or situation has the potential for interactivity. That is, the situated tasks, or simulated environments, tend to be far more interactive than simple discrete items. Interactivity might be valuable for assessments in which the examinee's ability to view information in context might be important. However, an increase in interactivity can potentially lead to a decrease in reliability, due to increased item dependence. Furthermore, interactive assessments might require additional training of item writers, as well as potentially requiring greater computers skills on the part of examinees. Finally, this type of innovative assessment is likely to be expensive to program and validate.

Media Inclusion

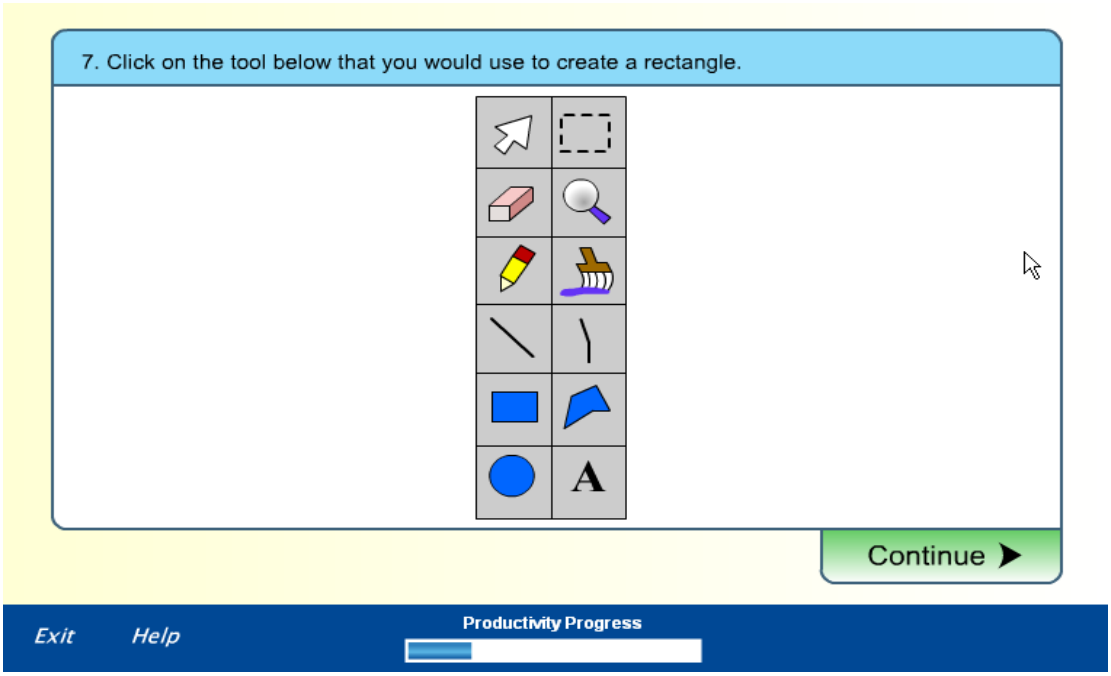
Another level of the taxonomy that is highly useful for the development of innovative items is that of media inclusion. The types of media that can be incorporated into the assessment include graphics, audio, video, and animation.

Graphics. Graphics are the most common type of non-text media included in computerized tests. Although paper-and-pencil tests can also include graphics, they lack the computer's facility for interactivity. On a computer, examinees may be able to rotate, resize, and zoom in or out of a scaled image, whether interacting with a graphical item stem or graphical response options. Examinees can respond by selecting one of a set of figures, by clicking directly on some part of a graphic (sometimes referred to as "hot spot" items), or by dragging icons to assemble a meaningful image. The Teacher Technology Skills assessment (Harmes et al., 2004) includes

items that provide reproduced elements of computer applications, such as the tool palette from a graphics-editing program. Examinees are asked to click on the area within the graphic that would be used to perform a specific action, such as creating a rectangle (see Figure 3).

The NCLEX exam contains items with elements such as a graphic of a human torso, and nurse candidates are asked to place the stethoscope (mouse pointer) on the area appropriate for performing part of a cardiac assessment (NCSBN, 2005). In the graphical modeling items presented by Bennett, Morley, and Quardt (1998), examinees respond by plotting points on a set of axes or grid, and then use either curve or line tools to connect the points. In a medical assessment, examinees are able to view high-resolution graphics, such as histopathology and other slides; the examinees are also able to pan across or zoom into these images to view them more closely (NBME, 2004). All of these examples incorporated visual elements that were highly content relevant. There are broad-ranging content applications for the use of graphics, and it is probably also true that graphics represent the easiest type of media to integrate into a CBT. Many software programs will easily integrate and store graphics in a variety of file formats, their file sizes tend to be small, and most examinees are comfortable with their inclusion.

Figure 3. Sample Hotspot Item
(© Florida Department of Education. Reproduced with permission).



Audio. Audio has been incorporated primarily into computerized tests of language skills and music, which are content areas that have traditionally assessed listening skills. Audio has been used in recent tests of English proficiency of non-native speakers (e.g., ACT, Inc., 1999; Arnold, 2003; Godwin, 1999; Nissan, 1999), and in tests of music listening skills (Perlman, Berger, & Tyler, 1993; Vispoel & Coffman, 1992; Vispoel, Wang, & Bleiler, 1997). However, there are certain advantages to administering audio in CBTs as compared to using paper-and-pencil along with cassette tapes. In computer-based tests, the sound quality is higher, and examinees can typically control the volume, timing, and possibly even frequency at which the clips are played (Parshall & Balizet, 2001).

Although there are many potential applications of audio that could increase the construct relevance of an assessment, there are also challenges to its use. It is critical that audio not be added in such a way as to create unnecessary or unfair disadvantages to hearing-impaired examinees. In addition, the logistical considerations of audio file type, storage requirements, and test security should be considered (Parshall & Balizet, 2001).

Video. Just as some conventional tests have long incorporated audio, so a few others have historically incorporated video. Videodiscs and videocassettes have been used in tests of business and interpersonal interactions, medical diagnosis and treatment, and aircraft operations (ACT, 1995; Bosman, Hoogenboom, & Walpot, 1994; Shea, Norcini, Baranowski, Langdon, & Popp, 1992; Taggart, 1995). Research examples of video-based items are reported in Bennett, Goodman, et al. (1997). Video incorporated within CBT has some technological advantages over these media, including greater reliability and examinee control of timing. A video-based test of conflict resolution skills was developed and validated by Olson-Buchanan, Drasgow, Moberg,

Mead, Keenan, and Donovan (1998). This test, which presents scenes of conflict in the workplace, also includes a level of interactivity in that an examinee's selection branches to the next video displayed. Video can be incorporated into a CBT as item stimulus material for text-based responses, and might also be included in the response options or actions.

Video appears to be a useful addition to an assessment when the construct relates to interpersonal communication or other aspects of human interaction. Furthermore, video has the capacity to display dynamic processes, such as moving pistons or a beating heart. In some of these dynamic applications either video or animation might be included. Video will be preferred when a great deal of "real world" detail is important. Additionally, the proliferation of digital cameras and video editing software will mean that video is easier to obtain in many instances.

One rationale for the inclusion of full-motion video, as opposed to just audio, is that it adds the non-verbal component of communication. However, many of the logistical issues that apply to audio apply to video as well. There are many possible file types, memory requirements are high, and test security could be problematic. In addition, video might be, on average, the most expensive media to incorporate in a CBT, depending on variables such as the use of professional actors or production studios.

Animation. Animation has the capacity to display dynamic processes, unlike paper, which is inherently a static medium, incapable of imparting more than a sense of movement to text or graphics. Although minimal use of animation has yet been made in testing, a few research examples can be found. Bennett, Goodman, et al. (1997), used a type of animation to display changes in nation boundaries over time, by displaying a series of static maps in quick succession. Examinees responded by identifying the particular static map that answered a question. Bennett, Goodman, et al. also noted that in many science and health related fields, professionals need to be able to read and listen to a variety of electronic instruments. They therefore developed a test item that included an animated heart monitor trace, a static electrocardiogram strip, and an audio file of the related heart sound. Finally, Martinez (1991) suggested figural response items that include stimuli such as time series processes (e.g., cell division). These examples illustrate content areas in which it might be important to assess examinees' understanding of dynamic processes. Animation has several advantages over video clips in certain applications, despite the relative popularity of the latter. Animation uses far less computer memory than video to store or display and, in some cases, it might be less expensive to produce. More substantively, animation is likely to be simpler, as it can focus the examinee on essential aspects of the movement more specifically than video is likely to do. For other applications, the "real-world" detail inherent in video might be essential.

The inclusion of media, whether graphics, audio, video, or animation, can strengthen the validity of an assessment when appropriately used. To insure high validity, the media files should be of sufficient quality. Attention should also be paid to considering the needs of visually-impaired or hearing-impaired examinees. A number of technological and logistical issues might also arise with the inclusion of media. A variety of file types might need to be supported, file sizes might be challenging for data storage and transmission, and test security concerns might arise if the media are highly memorable.

Response Action

We use the term response action to refer both to the types of input devices required as well as the physical action that an examinee makes to respond to an item or to complete a task. The most commonly used input devices in CBTs are the keyboard and the mouse. Examinees respond through the keyboard by typing numbers, characters, or more extended text. Examinees might use a mouse to respond to a multiple-choice item; to click on a graphic; or to access onscreen menus, exhibits, or other resources. They might need to drag icons to create or complete an image, or to drag text, numbers, or icons to indicate a correct sequence. As an alternative to dragging, some computerized exams ask examinees to click on an image, and then click a second time to mark the new location where the image should be placed.

For particular applications, use of input devices such as touch screens, light pens, joysticks, or trackballs might benefit measurement. For example, very young examinees might be assessed with less error using touch screens or light pens. Particular skills, such as those that are highly movement-oriented, might be better measured using joysticks or trackballs. Microphones and speech recognition software might be used to collect spoken responses to oral questions (Stone, 1998). Haptic devices (Gruber, 1998), which use force feedback to simulate touch in a 3-D environment, could greatly increase the real-world congruence of measurement in some arenas.

Moving away from the traditional presentation of discrete items, toward interaction within virtual environments, leads to a new dimension of response actions. Simulated environments and tools have long been used for training purposes (e.g., the Link flight trainer ca. 1930s; Aerospace Education Center, 1998). Although much of the current use is in aviation and medicine, other industries using simulated environments and tools include nuclear power plants, law enforcement, maritime transportation, and hazardous materials (GSE Systems 2005; Advanced Interactive Systems, 2005; California Maritime Academy, 2004; Industrial Scientific Corporation, 2005). As technology has developed, the level of sophistication and fidelity has increased to the extent that some of these devices are now being used for assessment purposes. The use of application-specific simulated devices results in a variety of potential new response actions.

It is important to remember, whether using the most common input devices such as the keyboard and mouse, or application-specific devices, that the response actions we require of examinees should be well within their abilities. If the computer-related skills are overly challenging, then we risk adding measurement error to the assessment. The response actions we require of examinees, and the input devices we ask them to use, should be appropriate for their computer skill level and should provide a meaningful and useful match to the construct under assessment.

Scoring Methods

Considerable work has been conducted in the area of expanding the automated scoring models that can be applied to assessments. Many of the important practical benefits of computerized testing require that scoring of the assessments be automated. Tests can be adapted to examinees only when item responses are instantly scored by the computer. Furthermore, score reports can be issued immediately after testing only when test scores are determined by the

computer. There are a number of questions that developers ought to consider when determining the type of scoring method the innovative assessment should use, as the scoring schema selected will affect test design and development. Examples of the kinds of questions the test developers might consider include: how to define a correct response, whether to score on single or multiple outcomes, and whether the multiple outcomes should include such elements as aesthetics and efficiency. Strategies for automated scoring of CBTs can be categorized into three broad areas: dichotomous, polytomous, and complex modeling. Each of these approaches has relative advantages and disadvantages.

The dichotomous approach to scoring involves collapsing the information provided by the examinee's response into a score of correct or incorrect. Advantages of the dichotomous approach to scoring are that it is easy to implement and that it is easy to justify to examinees and other test score consumers. One disadvantage of this approach is that some information might be lost. For example, if a situated task requires an examinee to complete several steps or take several actions, a simple dichotomous score of the final product ignores many elements of the examinee's response. Another disadvantage of dichotomous scoring is that its use might require the tasks to be constrained to fit this scoring model. That is, when designing for dichotomous scoring, the components of the assessment must be structured so that one option or outcome is clearly right and all others are clearly wrong. This constraining of the task, or pre-processing of the information, might be more or less awkward, depending upon the content area of the assessment. In the worst cases, the additional task constraints can result in a kind of construct-irrelevant variance that makes the task more confusing or difficult for the examinee, and thus results in measurement error.

Polytomous, or partial-credit, scoring represents an attempt to incorporate into the score some of the extra information that might be collected when an examinee responds to an item or completes a task. There are many ways in which polytomous scoring can be accomplished. In the case of selected response items, the various response options can be weighted for correctness, so that scores other than 0 or 1 are possible. With constructed response items, or in more complicated assessment structures, weighting could be applied to acceptable components of a response or to the individual steps within a task. Alternatively, polytomous scoring might involve a careful, considered effort to break down the task into very discrete pieces, collect data from the performance record and then evaluate and combine scores on the pieces (see Harmes et al., 2004, for a relatively simple example.) The choice of a polytomous scoring model might require that decisions be made regarding the number of criteria to evaluate, along with the relative weight of each criterion. For example, an IT certification exam might include simulated software that examinees use to complete a task. The task must then be designed to allow for various types of responses and the collection of the additional process information, such as time or number of steps taken. The score for the task might include both the correctness of the examinee's final response, as well as the efficiency of the process taken. The primary advantage of the polytomous approach is the increase in information over dichotomous scoring, while the primary disadvantage is the greater demand for programming, calculating and explaining of polytomous scoring**.**

As assessments become more innovative, more elaborate scoring methods might be needed. Highly integrated situated tasks and simulated environments might include a larger set of

acceptable variations in examinee responses than do constructed-response or simple situated tasks. These types of assessment structures might be difficult to evaluate using a strictly logical framework. Some of the complex modeling techniques, such as automated essay scoring methods, focus on the examinee's final product, while others focus on the examinee's process. A common approach to initiating an automated scoring system involves identification and evaluation of the salient, measurable elements of the product or performance, followed by the development of a model for combining these elements into a score. In contrast to polytomous scoring, complex modeling approaches attempt to emulate or model the process that would be taken by human raters to score, or by expert examinees to complete the task.

More sophisticated scoring models have the potential to capture greater information for the examinee's response. However, that potential greater information at the cost of greater effort in the development, programming, and validation of the scoring model. One risk in developing a more elaborate scoring model is the possibility of developing an incomplete or biased scoring model. In additional, some polytomous approaches to scoring can be difficult to explain to examinees and other stakeholders.

Using the Taxonomy to Design Item Templates

Introduction to Templates

Templates are a structured means of collecting and storing item information, such as specifications and elements. Templates for innovative items incorporate the ideas of storyboards from the field of instructional design, along with item frames or forms (Downing, 2006). Some testing programs are currently using templates to improve the structure, efficiency, and security of the innovative item writing process.

Once an innovative item type has been developed, a template can be designed to provide structure for item writers. Specifically, a template can help guide and constrain item writers who have been charged with writing items using the innovative item type. A highly detailed template can also increase the efficiency of programming and media production. Finally, some templates can help provide security through the use of item variants as alternate or substitute items.

Although there are many types of templates in use, a primary aspect of many templates is the provision of specified database fields related to the item type. Most item types will include fields such as the stem, key, author, and reference; the template that has been designed for a specific item type will include additional relevant fields. For example, a hot spot item will include a graphic (whether linked or embedded) and an identification of the correct area on the graphic. An example of a template with these elements is provided in Figure 4. Templates may include screen design and layout elements. In the hotspot item example, this might include additional requirements such as specification of all areas that should be available for selection by examinees. This might involve providing the graphic within the template and having the item writer mark on the graphic, in addition to naming the areas and classifying them as correct or incorrect. Figure 5 displays an example of a template for a different item type, this one including layout and design elements for an item that uses a simulated software program as part of an IT exam.

Figure 4. Sample Template Illustrating Database Fields for Item Information

Item ID #:	Keywords:
Author:	Reviewer:
Instructions:	Reference:
Prompt:	
Graphic file name:	
Correct area(s):	
Incorrect area(s):	

Figure 5. Sample Template Illustrating Screen Layout Elements

Task Description:	
Instructions:	Simulated software program:
Navigation Options:	

Test Development Concerns and the Taxonomy

When innovative item types are included in an exam program there is an inevitable impact. The nature of that impact, as well as the extent of it, will be based on the specific taxonomy levels included as components of the innovative item types developed. The seven levels of the taxonomy will differentially impact test development concerns such as the exam program construct, psychometrics, programming needs, examinee computer skills, and cost. A careful

consideration of the taxonomy levels in light of these test development issues might help test developers as they make decisions about the design of a new innovative item type.

Table 1 provides a brief summary of some of the effects that innovation might have on aspects of an exam program. In general, the more "innovative" an item type is, the greater the effect is likely to be. For example, an item type that incorporates interactivity to a modest extent is likely to need some additional programming; a highly interactive assessment will usually require far greater programming effort. This will also impact the amount and type of information that must be specified within an item template.

This table could be helpful as the types of innovative items for an exam program are being considered. An illustration, using the taxonomy level of response action, can be considered. Assume that a given exam program had a construct-relevant reason for incorporating innovative response action, but that no alternative input device is needed. Additional programming is likely to be needed; this will increase the initial cost of development. However, it might not increase the ongoing costs greatly. A further consideration is that this use of an innovative response action might mean that the assessment requires higher levels of computer skills on the part of the examinees. Finally, it will be important to consider whether the response action provides good fidelity in some "real world" settings, but an actual mismatch in others, as this could result in construct irrelevant variance for some examinees. All of these issues should be evaluated in conjunction, to determine whether or not this specific type of innovation is appropriate for this exam program.

Table 1. The Impact of Taxonomy Levels on Test Development Concerns

Taxonomy Level	***Exam program construct***	***Psychometrics***	***Programming needs***	***Examinees' computer skills***	***Cost***
Assessment structure	Potential for improved measurement	Less is known about the psychometric functioning of most innovative item types	Additional programming may be needed	Higher computer skills may be required (depending on the type of assessment structure).	Initial or ongoing costs may increase
Complexity	Potential for improved measurement	Additional analyses may be needed	Additional programming may be needed	Higher computer skills may be required	Initial or ongoing costs may increase
Fidelity	Potential for improved measurement	Potential for construct irrelevant variance from distracting elements	Additional programming may be needed (depending on type of fidelity)	Higher computer skills may be required (depending on type of fidelity)	Initial or ongoing costs may increase (depending on type of fidelity)
Interactivity	Potential for improved measurement	Additional analyses may be needed	Additional programming may be needed	Higher computer skills may be required	Initial or ongoing costs may increase
Response action	Potential for improved measurement	Potential for construct irrelevant variance if response action is dissimilar to real world	Additional programming may be needed	Higher computer skills may be required	Initial or ongoing costs may increase
Media inclusion	Potential for improved measurement	Additional analyses may be needed (depending on type of media)	Development and programming likely to increase	Higher computer skills may be required (depending on type of media)	Initial or ongoing costs may increase
Scoring algorithm	Potential for improved measurement	Additional analyses may be needed	No apparent impact	No apparent impact	Initial or ongoing costs may increase

The Taxonomy and Template Design

The consideration of the taxonomy levels and their impact on test development concerns can be conducted during the process of drafting, researching, and refining an innovative item type. Once the item type has been fully designed, a template for the item type can also be designed. The template will be specific to the item type and, potentially, to the exam program as well. It should include database fields for every piece of information that ought to be provided for every item based on that item type. Depending upon the type of innovative item, the template may include typical information that would appear in an item banking program, along with fields for specifying file names, screen placement, actions or reactions related to specific elements, etc.

Conclusions

In this paper we have presented a new taxonomy for assessment innovations based on the seven levels of assessment structure, complexity, fidelity, interactivity, response action, media inclusion, and scoring algorithm. Each dimension of the taxonomy has the potential for both costs and benefits. Illustrations of the taxonomy levels were provided through a number of examples of innovative assessments. These examples indicate the range of innovative assessment developments that are already in use, as well as giving hints of the kinds of assessments we might expect to see in the future. In the best cases the innovations provide true measurement improvement.

We have suggested the utility of designing templates for innovative item types. These templates might be developed as the final step in the design of new item types. They would codify the decisions made after test development staff have fully considered the goals and requirements of the exam program, as well as evaluating the impact of the new item types on important test development concerns. Once these design decisions have been made, templates can provide a useful structure to item writing. Templates can guide and constraint the item writing process; they can further support the efficiency and security of an exam program by providing a framework for creating item variants.

We hope these materials prove useful to other test developers and researchers interested in expanding our forms of measurement through the use of innovative item types.

References

ACT, Inc. (1995). *Work Keys.* Iowa City, IA: Author.

ACT, Inc. (1999). *Technical manual for the ESL exam.* Iowa City, IA: Author.

AICPA. (2004, February). *AICPA, NASBA, and Prometric successfully pilot computer-based exam for CPAs.* Retrieved April 9, 2005, from http://www.aicpa.org/download/news/2004_02_02.pdf

AICPA. (2003*). Uniform CPA Examination tutorial*. Retrieved May 9, 2005, from http://www.cpa-exam.org/tutorial/index.html

Advanced Interactive Systems. (2005). *PRISim video-based judgment training simulator*. Retrieved July 2, 2005, from http://www.ais-sim.com/prisim.htm

Arnold, A. (2003). The TOEFL CBT. *Language Testing, 20*, 111-123.

Aerospace Education Center. (1998). *Link trainer*. Retrieved July 2, 2005, from http://www.aerospaced.org/permart/linkt.html

Bennett, R. E., Goodman, M., Hessinger, J., Ligget, J., Marshall, G., Kahn, H., & Zack, J. (1997). *Using multimedia in large-scale computer-based testing programs.* (Research Rep. No. RR-97-3). Princeton, NJ: Educational Testing Service.

Bennett, R. E., Morley, M., & Quardt, D. (1998, April). *Three response types for broadening the conception of mathematical problem solving in computerized-adaptive tests*. Paper presented at the annual meeting of the National Council of Measurement in Education, San Diego, CA.

Bosman, F., Hoogenboom, J., & Walpot, G. (1994). An interactive video test for pharmaceutical chemist's assistants. *Computers in Human Behavior, 10*, 51-62.

Braun, H. (1994). Assessing technology in assessment. In Baker, E. A., & O'Neil, H. F. (Eds.), *Technology assessment in education and training* (pp. 231-246). Hillsdale, NJ: Lawrence Erlbaum Associates.

Breland, H.M. (1998, April). *Writing assessment through automated editing*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego.

California Maritime Academy. (2004). *Simulator training.* Retrieved July 2, 2005, from http://www.csum.edu/academics/simulatortraining.asp

Davey, T., Godwin, J., & Mittelholtz, D. (1997). Developing and scoring an innovative computerized writing assessment. *Journal of Educational Measurement, 34*, 21-41.

Dillon, G. F., Boulet, J. R., Hawkins, R. E., & Swanson, D. B. (2004 October). Simulations in the United States Medical Licensing Examination (USMLE). *Quality and Safety in Health Care, 13*, i41-i45.

Downing, S. M. (2006). Selected-response time formats in test development. In Downing. S.M, & Haladyna, T. M. (Eds.), *Handbook of test development* (pp 287-301). Mahwah, NJ: Lawrence Erlbaum.

French, A., & Godwin, J. (1996, April). *Using multimedia technology to create innovative items*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY.

Gruber, J. S. (1998, October). [Interview with James Kramer, head of Virtual Technologies, Inc.] Gropethink. *Wired*, pp. 168-169.

Godwin, J. (1999, April). *Designing the ACT ESL Listening Test*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

GSE Systems. (2005). *Nuclear simulation*. Retrieved July 2, 2005, from http://www.gses.com/powernuclear.html

Harmes, J. C. & Parshall, C. G., (2005). *Situated tasks and simulated environments: A look into the future for innovative computerized assessment.* Paper presented at the annual meeting of the Florida Educational Research Association. Miami, FL.

Harmes, J. C. & Parshall, C. G. (2000). *An iterative process for computerized test development: Integrating usability methods.* Paper presented at the annual meeting of the Florida Educational Research Association. Tallahassee, FL.

Harmes, J. C., Parshall, C. G., Rendina-Gobioff, G., Jones, P. K., Githens, M., & Dennard, A. (2004, November). *Integrating usability methods into the CBT development process: Case study of a technology literacy assessment*. Paper presented at the annual meeting of the Florida Educational Research Association, Tampa, FL.

Heffner, T. S., Knapp, D. J., Rosenthal, D. (2004, November). *Pushing the bounds of testing technology: Designing an innovative, affordable testing program for multiple occupations.* Presentation at the annual meeting of the National Organization for Competency Assurance.

Industrial Science Corporation. (2005). *Metering/confined space simulator training*. Retrieved July 2, 2005, from http://www.indsci.com/serv_train_meter.asp

Koch, D. A. (1993). Testing goes graphical. *Journal of Interactive Instruction Development*, 5, 14-21.

Longridge, T., Burki-Cohen, J., Go, T. H., & Kendra, A. J. (2001, March). Simulator fidelity considerations for training and evaluation of today's airline pilots. *Proceedings of the 11th Annual Symposium on Aviation Psychology, Columbus, OH.* Retrieved April 20, 2005, from http://www.volpe.dot.gov/opsad/docs/isap-2001.doc

Luecht, R. M. & Clauser, B. E. (2002). Test models for complex CBT. In C. N. Mills, M. T. Potenza, J. J. Fremer, & W. C. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 67-88). Mahwah, NJ: Lawrence Erlbaum Associates.

National Council of State Boards of Nursing. (2005 March). *Fast facts about alternate item formats and the NCLEX examination.* Retrieved April 20, 2005, from http://www.ncsbn.org/pdfs/01_08_04_Alt_Itm.pdf

Martinez, M. E. (1991). A comparison of multiple choice and constructed figural response items. *Journal of Educational Measurement, 28,* 131-145.

Martinez, M. E. (1993). Item formats and mental abilities in biology assessment. *Journal of Computers in Mathematics and Science Teaching, 12,* 289-301.

National Board of Medical Examiners. (2004 Fall/Winter). Continuing developments in computer-based testing. *NBME Examiner*. Retrieved May 9, 2005, from http://www.nbme.org/Examiners/fallwinter2004/news2.asp

Olson-Buchanan, J. B., Drasgow, F., Moberg, P. J., Mead, A. D., Keenan, P.A., & Donovan, M.A. (1998). Interactive video assessment of conflict resolution skills. *Personnel Psychology*, 51, 1-24.

O'Neill, K., & Folk, V. (1996, April). *Innovative CBT item formats in a teacher licensing program*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY.

Parshall, C. G., & Balizet, S. (2001). Audio computer-based tests (CBTs): An initial framework for the use of sound in computerized tests. *Educational Measurement: Issues and Practice,20,* 5-15.

Parshall, C. G., Davey, T., & Pashley, P. (2000). Innovative item types for computerized testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice.* (pp.129-148). Norwell, MA: Kluwer Academic Publishers.

Parshall, C. G. & Harmes, J. C. (2005, February). *Tools for improving the CBT user-interface: Paper prototyping, expert review, and user testing.* Workshop presented at the annual meeting of the Association of Test Publishers, Scottsdale, AZ.

Parshall, C. G., Stewart, R, & Ritter, J. (1996, April). *Innovations: Sound, graphics, and alternative response modes.* Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Perlman, M., Berger, K., & Tyler, L. (1993). *An application of multimedia software to standardized testing in music.* (Research Rep. No. 93-36). Princeton, NJ: Educational Testing Service.

Shea, J. A., Norcini, J. J., Baranowski, R. A., Langdon, L. O., & Popp, R. L. (1992). A comparison of video and print formats in the assessment of skill in interpreting cardiovascular motion studies. *Evaluation and the Health Professions, 15,* 325-340.

Stone, B. (1998, March). Focus on technology: Are you talking to me? *Newsweek,* 85-86.

Taggart, W. R. (1995). Certifying pilots: Implications for medicine and for the future. In E. L. Mancall & P. G. Bashook (Eds.), *Assessing clinical reasoning: The oral examination and alternative methods* (pp. 175-182). Evanston, IL: American Board of Medical Specialties.

Tsai, T-C., Harasym, P. H., Nijssen-Jordan, C., Jennett, P., & Powell, G. (2003 November). Simulation in assessment: The quality of a simulation examination using a high-fidelity child manikin. *Medical Education, 37(s1),* 72.

Vispoel, W. P., & Coffman, D. (1992). Computerized adaptive testing of music-related skills. *Bulletin of the Council for Research in Music Education, 112,* 29-49.

Vispoel, W. P., Wang, T., & Bleiler, T. (1997). Computerized adaptive and fixed-item testing of music listening skill: A comparison of efficiency, precision, and concurrent validity. *Journal of Educational Measurement, 34,* 43-63.

Zenisky, A. L. & Sireci, S. G. (2002). Technological innovations in large-scale assessment. *Applied Measurement in Education, 15(4),* 337-62.