Improving the Quality of Innovative Item Types: Four Tasks for Design and Development

Cynthia G. Parshall, Ph.D.
Measurement Consultant

J. Christine Harmes, Ph.D.
James Madison University

Abstract

Many exam programs have begun to include innovative item types in their operational assessments. While innovative item types appear to have great promise for expanding measurement, there can also be genuine challenges to their successful implementation. In this paper we present a set of four activities that can be beneficially incorporated into the design and development of innovative item types. These tasks are: template design, item writing guidelines, item writer training, and usability studies. When these four tasks are fully incorporated in the test development process then the potential for improved measurement through innovative item types is much greater.

Introduction

In recent years the proliferation of computer-based tests (CBTs) has been accompanied by the development and operational use of many new item types (Parshall, Spray, Kalohn & Davey, 2002). Innovative item types, defined broadly, are those items in a CBT that make use of features and functions of the computer to do things not easily done in traditional paper-and-pencil assessments. Innovations could address: 1) the item format or assessment structure, 2) the response action, 3) any media included in the item, 4) any interactivity provided by the item, 5) the complexity of the item, 6) the fidelity of the item to the real world, and 7) the scoring method used for the item (Parshall, Harmes, Davey & Pashley, in press).

The appeal of including innovative items is usually based primarily on the possibility of expanding the measurement of the construct. The better the match between the form of the item and the construct, the smaller the leap between scores on an item and the inferences made from those scores. To the extent that an innovative item allows a testing program to have more direct measurement, then measurement can be improved. An additional benefit of many innovative items is that they can reduce the potential for examinees to guess the correct response. When measuring many content areas with text-based multiple-choice items, pre-processing has to occur. This pre-processing includes narrowing the range of choices from a very large number that might be encountered in a realistic setting to a prescribed 4 or 5 options. Reducing the opportunity for guessing is another way in which innovative items can help improve measurement. For example, a hot spot item that allows an examinee to click on any area in the picture can reduce the likelihood that a correct choice of one out of four selectable areas happened by chance.

Incorporating innovations into a testing program should be done thoughtfully. Adding innovative items simply to "keep up" with technological advances is not recommended. Rather, innovations should be purposefully selected so that they serve the purpose of expanding or improving measurement. It is also important to note that adding technology brings with it the potential for unexpected modifications of the construct being measured. For example, there is a risk of introducing construct-irrelevant variance through a poor user interface or unclear response action requirements. Furthermore, it is possible that presenting stimuli through multimedia instead of reading text may actually alter the construct being measured. On the other hand, if the technology or other innovation is incorporated purposefully, then any changes to the construct should be positive.

Parshall, C. G., & Harmes, J. C. (2009). Improving the Quality of Innovative Item Types: Four Tasks for Design and Development. *Journal of Applied Testing Technology 10(1)*.

In order to ensure the best possible results from the addition of innovative item items, we recommend a careful approach. In most cases, when an innovative item type is being added to an exam program, it is a *new* item type for that program. Much less psychometric information about these new item types is available, meaning that test developers and item writers may have less expertise to draw on in producing high quality items. For these reasons, we believe that a thorough *design stage* for the new innovative item type is useful and important. There are also tasks that can be used to strengthen the high-quality *development* of new item types.

In this paper we introduce a set of four activities that can be beneficially incorporated into the design and development of innovative item types. These tasks are: template design, item writing guidelines, item writer training, and usability studies. The last of these activities, usability studies, may be new to most test developers, although it has an extensive history of successful use in the field of Human-Computer Interaction (HCI). The remaining tasks are all used in the development of traditional items. However, it will often be beneficial to modify or expand these tasks when innovative items are being used, in order to appropriately address all of the new elements that may be present.

A Model for Designing Innovative Item Types

One model for the design of innovative item types involves a 6-step process that includes several rounds of review and revision (Parshall & Harmes, 2008a). The steps in this model are: 1. analyze the exam program's construct needs, 2. select specific innovations for consideration, 3. design initial prototypes for internal discussion, 4. iteratively refine the item type designs, 5. conduct a pilot test of the innovative item types, and 6. produce final materials. The tasks presented in this paper are conducted as part of Step 4, when this full model is implemented. The model is provided in Figure 1.

The first step of this model, *analyze the exam program's construct needs*, consists of a thoughtful consideration of the exam program's current measurement successes as well as an identification of weaker or even missing areas. Step 2, *select specific innovations for consideration*, turns the focus on approaches to innovative item types that may be used to address those construct needs. In Step 3 the test developers, in collaboration with subject matter experts (SMEs), begin to define the new item types for the exam program, based on the selected innovations. Once a preliminary item type design has been specified, then an *initial prototype is designed* for preliminary consideration by internal exam program stakeholders. This initial review phase is likely to result in some modifications to the item type, prior to Step 4.

Step 4, *iteratively refine the item type designs*, is the most extensive step in this model, in part due to its iterative nature, and is the focus of this paper. Three related set of activities are undertaken in a series of interconnected rounds or iterations. The three activities are: develop initial item writing materials and sample items, conduct usability testing on the sample item types, and conduct extensive stakeholder reviews. Within each iteration, feedback is input and revisions are attempted in order to arrive at an improved item type design. It is anticipated that most proposed item type designs will need to proceed iteratively through all the Step 4 activities in several successive cycles to be fully defined and appropriately specified.

Step 5 of the model is *conduct a pilot study*. The pilot effort should occur after the item type designs have been iteratively revised and have reached a satisfactory level of quality. Pilot testing of the new item types should include a test of all exam program systems, such as: item banking, test publishing, test delivery and administration, examinee response capturing, item analysis, and test scoring. Once item types have been successfully pilot tested, they are ready for operational implementation. In the final step, Step 6, *all relevant exam program materials and documentation* are updated to include the new item types.

This complete 6-step model for item type design is intended to help exam programs add innovative items that are of high measurement quality, logistically practical, and acceptably affordable. Nevertheless, an exam

program might elect to undertake only some of the activities, iterations, and steps in this model. Each of the four test development tasks addressed in this paper has the potential to improve the design and development of new item types. Each of these tasks (template design, item writing guidelines, item writer training, and usability studies) is described in some detail next.

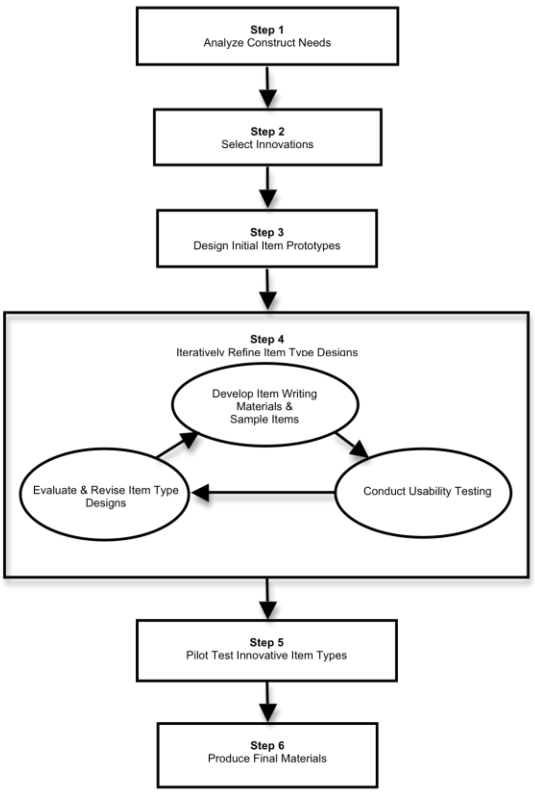

Figure 1. *Process for Innovative Item Type Design (from Parshall & Harmes, 2008a)*

Template Design

Template use in assessment stems from earlier work on item forms or shells (Roid & Haladyna, 1982). Templates are a structured means of collecting and sometimes storing data related to an item type. That data may be used in the development of individual items. Templates are sometimes used with traditional items, but are of particular interest with innovative item types. For innovative items, templates can be used in both the design stage and the development stage.

Item type templates can serve as a framework during design and planning stages. Templates can reflect visual components of an item type, including layout and other aspects of screen design. In addition, templates can be used as the exam program considers questions such as: *how many clickable areas can be specified for a hot spot item created from the template*? Another question might be: *which of the item components will be changeable or definable by the item writer*? (e.g., *will there be a common prompt for items developed from the template*?). Once a template has been designed and specified, one or more prototypes can be developed from the template. A prototype is an instantiation of a template; it is a completed item or task that is built using the framework of the template. As Figure 1 above shows, item prototypes are initially created and tested in Step 3 of the design model, when they can be used for internal evaluation of the new item type. Prototypes are also useful during development, to give item writers clear examples and models of the item's features and elements that need to be specified.

At the test development stage templates are highly useful as an item writing tool. While traditional paper-and-pencil exam programs may use item templates as a support to the item writing process, their use is particularly important for innovative items. Each newly written innovative item may need specific, detailed information about graphics, sound, video, or other critical content specifications. Without templates the risk of receiving incomplete information from the item writers becomes substantial and can result in logistical challenges and multiple follow-up requests of the item writers (Parshall & Becker, 2008).

Template Use in Design

Initially, templates can be used by the development team as a planning device during the stage of new item type design. Template design may address item components and/or item screen design. When developing a template for a new item type, the development team must first decide on a detailed set of specifications for each item type. These specifications are then translated into a template. Once these elements have been comprehensively detailed, the development team then needs to specify which components will remain static across all items developed from the template, and which components will be created by the item writers for each actual item or task. For example, a multiple choice with graphics item type may retain as fixed elements of the template the same layout and the same item prompt (e.g., "*Choose the picture that best depicts the scenario described below"*). When a template such as this is used by item writers they would keep the fixed elements while creating the scenario and response options, as well as providing a description of the graphic to be created. In some cases it is particularly useful to design various levels of sub-templates, each built upon the initial template. Generally, as more specific sub-templates are created, more item or task components are fixed (i.e., fewer components will be designed or created by item writers).

Depending upon the item type, an important part of the template planning process may be screen design and layout. If, for example, the new item type is a multiple choice item with graphics, the template developers may establish a standard layout for this item type; it may be decided that in all instances of this item type the text will appear on the left half of the screen and the graphic will appear in the right half of the screen.

The initial template is likely to go through a series of iterations as the item type design is being refined by the full team. This process often includes the development and review of several low-level prototypes for the template. These low-level prototypes may be simple representations of the item in a slideware program such as *PowerPoint*. When the design phase is complete, the template should be a functional item writing tool that can support ongoing item writing efforts.

Template Use in Development

In some instances, item writers may use the template as a guide and set of specifications so that they attend to all necessary elements when writing the items. In other instances the template may be the actual data collection device itself (i.e., be integrated with an existing item banking program or be developed into a customized tool for item banking). The exact method in which templates are applied will depend on the structure of the exam program and the item banking software in use. In any case, the templates can be used to improve item structure, production efficiency, and exam security. Each of these benefits will be described next.

Item Structure

Templates can help improve *item structure* by standardizing the way in which each new item format will be constructed and presented. That is, item writers are given a template for each item type in the exam program. A specific template might include database entry fields to be completed (i.e., the exact data to be collected, in the required format) and guidelines for the item writers. The guidelines, depending on the item type, might address such elements as the length of any video clips, the number of selectable areas for a hot spot item, the screen space available for a graphic, etc. Providing these details within the template can help

item writers make appropriate decisions as they create the items. Furthermore, these fully specified templates can be thought of as a menu of choices for an item writer. As the item writer begins to address a particular area of the test blueprint, he or she can select the specific template that is optimal for addressing the targeted content material.

An example of a very basic template is given in Figure 1. This type of template simply specifies the fields the item writers must complete for a hot spot item type, and serves as the data collection device.

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

Figure 1. Basic hot spot item template with database fields.

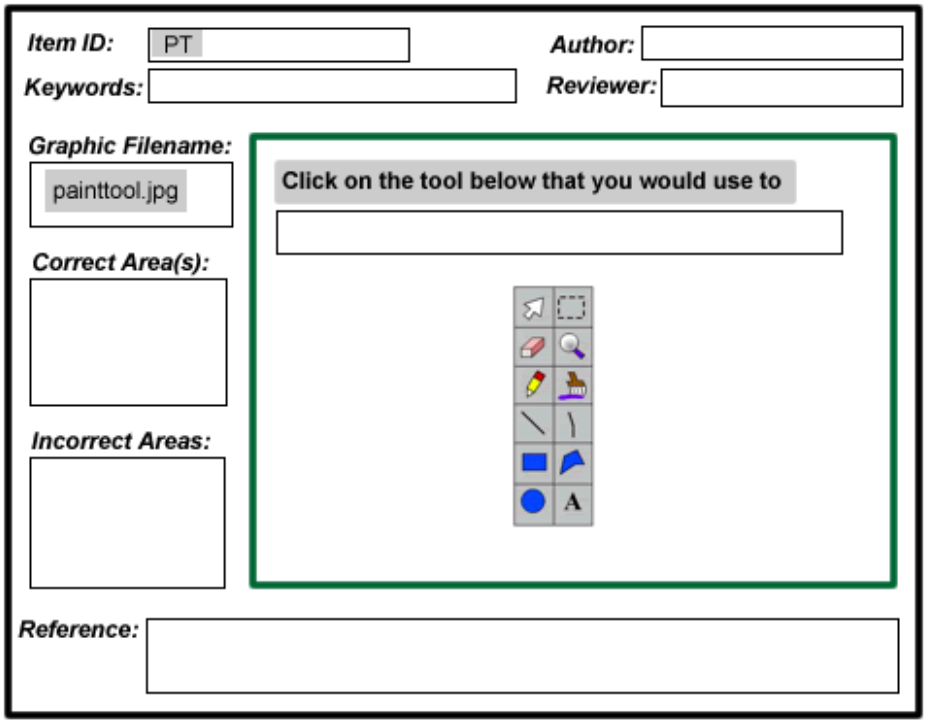

Figure 2. Sub-template for a hot spot item.

Figure 2 uses a different visual display for many of the same database elements. In addition, this example template includes a visual representation of how the item would appear on the screen. Figure 2 further illustrates various aspects of template use in that it is a sub-template of the basic template presented in Figure 1. In this sub-template, some elements have been specified to remain consistent across items created from the template, while others remain available for modification or supplementation. For example, with the sub-template in Figure 2, item writers would complete the information such as the remainder of the stem, keywords, and a reference. They would also type in a text description of the correct and incorrect areas.

A prototype item generated from this sub-template is provided in Figure 3.

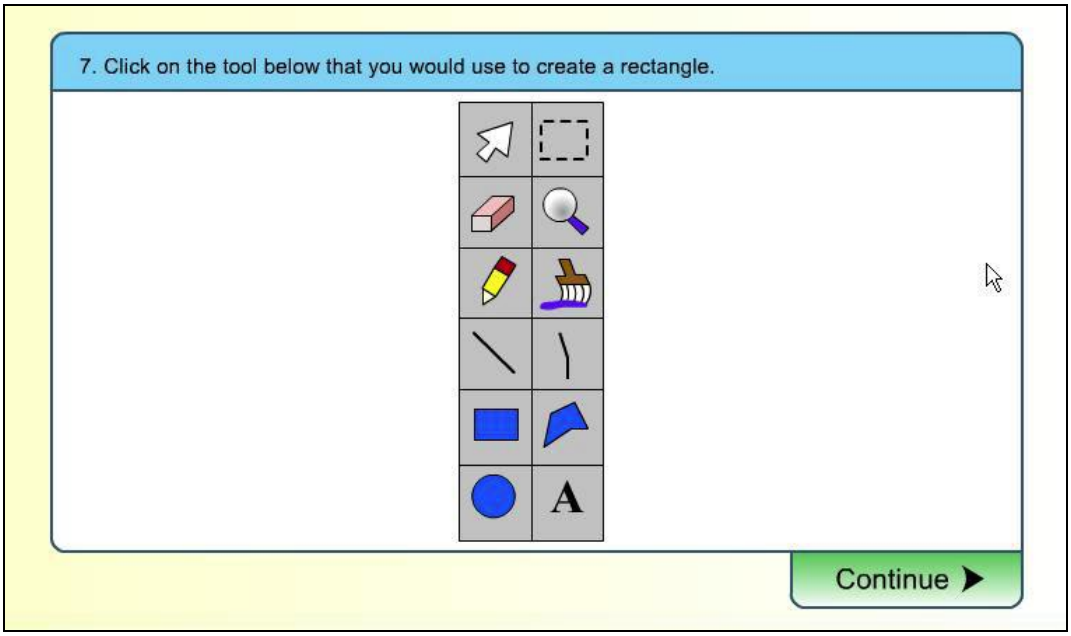

Figure 3. Prototype hot spot item created from sub-template.

A basic template for a multiple-choice with audio item may be quite similar to Figure 1, replacing the field for "Graphic file name" with "Audio file name". Instead of fields for correct and incorrect areas, there could be fields for voice actor instructions, script, and actor specifications (such as age or gender). From this general template, a sub-template could be created for variations on this particular audio item design. Such a sub-template could provide an additional level of specification, further constraining the item writer's task, while helping to ensure consistency across similar instantiations of this task type. This could be done by creating a standard set of instructions for items developed from the template, a common prompt, standard placement of elements on the screen, or even a common stimulus audio file.

Regardless of the number of template levels that are created, their structure should help streamline the item writing and development process, as item writers' tasks are constrained and their responsibilities for making design decisions are reduced. This intentional focus both clarifies and simplifies the item writer's task.

Production Efficiency

In a related fashion, *production efficiency* may also be improved, as item writers can be tasked with filling in components of a template, instead of creating an entirely new item concept each time. For example, the template in Figure 2 might yield one item that requires examinees to choose the tool to use when entering text, while another item based on the same template might require choosing the tool for use in selecting a

portion of a picture. A template specifies elements that will be included in each instantiation of the item (such as general instructions, or a standard set of response options, etc.) and empty fields where item writers need to create content. When these items are produced, many of the same elements (such as graphics, instructions, simulated software components, etc.) can be used across the various instantiations of each item template.

Another example of template use for production efficiency can be envisioned within a test of communication skills. In this case, an item that includes audio in the stem might have a standardized set of instructions for playing the audio, and a standard stem (e.g., "*Which of the following is the most appropriate response?"*). The item writer would then only have to create the script for the audio clip in the stem and the response options. Additional constraints could be specified for the item writers. For example, *the audio clip cannot be any longer than one minute, no more than two actors may be used*, etc.

Exam Security

Finally, *exam security* may also be improved, since templates can be used as tools for quickly developing different versions of an item. Depending upon the level of specificity to which the template has been designed, additional versions of an item might be created by substituting one descriptive phrase within the stem for another (such as a name or location), or by substituting one media element for another (such as a male voice for a female voice in an audio clip). For example, one instantiation of the hot spot template in Figure 2 might ask examinees to create a circle. The graphic used in this template contains 12 buttons, so there are theoretically 12 different items that could be created from this sub-template. Thus, very little effort would be required on the part of item writers to make an additional item, but examinees trying to disclose this item, or pass it on to others, might be thwarted.

Item Writing Guidelines

In addition to the development of item templates, expanded item writing guidelines should be produced for each new item type. Item writing guidelines have been used in traditional paper-and-pencil exam programs for many years. These standard item writing guidelines have a well established effectiveness in improving and maintaining high quality traditional items. In fact, one obvious reason for the success most standardized exam programs have with multiple choice items is that clear guidelines, based on decades of experience, are available for this item type (see for example Haladyna, 1996).

However, most innovative item types include new elements that are not fully addressed in the existing item writing guidelines. For this reason, Step 4 of the model for innovative item type design includes the development of expanded item writing guidelines.

Depending on the specific innovation, new elements that need to be addressed may include technology concerns, cognitive elements, psychometric issues, or examinee instructions. Specific decisions may need to be made about each of these facets; in many cases the decisions will then be formalized as guidelines for the item writers. (While item writing guidelines for innovative items have not been broadly disseminated in the measurement literature, some exam programs may have developed item writing guidelines and may be using them internally, even though they have not been published or presented externally.)

It seems quite reasonable that the process of developing new guidelines would begin with a consideration of features specific to the new item type. These features should be analyzed, in light of the exam's construct, with an eye towards possible constraints or specifications that may be useful. This analysis can lead to decisions about how the item features should be addressed in the implementation of the item type. For example, innovative items that include graphics may produce more consistent results when item writing guidelines specify requirements for the appearance of the images as well as aspects of how the images should be incorporated with other item content.

In the next few sections this process is illustrated, using two specific types of innovative items, *audio* items and *hot spot* items, as examples. The decisions and guidelines for other item types would be different, but the consideration of technical, construct, and measurement issues should be similar.

The Audio Item Type

Audio has traditionally been included within assessments of music and language, even in paper-and-pencil testing. As CBTs have made the inclusion of audio a relatively simple matter, exam programs in other content areas have also begun to use sound (Parshall & Balizet, 2001). These newer applications of audio most commonly use speech sounds, either in tests of oral communications or in voiceover accommodations. However, non-speech sounds, such as those produced by some types of medical equipment, can also be included in exams.

A variety of issues arise when audio is included in an assessment. Decisions may need to be made about the number of speakers used in any given audio file, the maximum allowed length for an audio file, the type of information that should be conveyed through audio, the level of examinee control over the sound files that should be provided in the CBT software, and the appropriate level of realism that should be reflected in the sound files. Each of these decisions may have implications for item writing which should then be addressed through the specification of clear item writing guidelines.

Number of Speakers

Many exam programs that use speech sounds have noted that examinees can become confused if too many different speakers are used. For this reason, one decision that is often made for audio items is that no more than two speakers should be used in a single audio file. A further, related, decision that might be made is that when two speakers are used, these should include one male voice and one female voice. This specification can help examinees to more easily keep track of who is speaking at each point in a dialogue.

Once decisions such as these are made they can be codified as item writing guidelines. An item writer, charged with writing items that include relevant speech scripts, would incorporate these constraints into the item writing process.

Length of Audio Files

Another area in which guidelines for audio items could be useful is that of the maximum time or length of each sound file. Decisions about the appropriate maximum length of an audio file are likely to vary across exam programs. For example, in language listening assessments some fairly lengthy audio prompts might be desirable, perhaps several minutes in length. In fact, this type of assessment might use several specific item types, addressing different aspects of the language listening construct with varying time limit guidelines. However, most content areas would be better served by restricting audio files to much shorter limits, perhaps as little as 15-30 seconds. Short audio files such as these are often appropriate in tests of other constructs simply because of the demands that listening places on short-term memory.

Type of Information

Evidence indicates that sound is processed in cognitively different ways from visual information (Ballas, 1994). Sound differs from visuals in that it is dynamic and ephemeral; in other words, it changes over time and then it is gone (Gaver, 1989). Furthermore, as noted above, audio in an item will often place greater demands on the examinee's short-term memory capacity than text generally does.

For all these reasons, it is important to consider the types of item information that should be communicated through sound. For example, interpersonal communication may be more effectively conveyed in spoken

dialogue, while complex detailed information may be better communicated through written text. The decisions about the type of audio information to be included in a given exam program should be closely tied to the construct goals of the assessment.

Examinee Control Over the Audio

Innovative items with audio frequently provide the sound in the item stems, typically by including a "Play" button. The examinee is expected to click the button and listen to the sound file, prior to responding to the item. The item itself may be a traditional item type such as the multiple choice.

In cassette- or CD-based tests with audio, rigid control is usually maintained over the timing of the audio prompts, as well as the number of times each audio is played. These controls are imposed partially due to technological restrictions of cassettes and CDs and to administrative restrictions necessitated by group administration of the audio. A CBT, on the other hand, can easily allow each examinee individual control over the timing of the audio file. CBT functions can also enable examinees to re-play the audio as many times as they wish.

While this difference in functionality can be technology-driven, it also poses clear measurement implications. For example, does the meaning of the construct change when examinees are allowed to re-play an audio file an unlimited number of times? For certain content areas (including many aspects of music and language), this free access might well render the items overly easy. These assessment areas often restrict audio to a maximum of two "repeat plays". In other applications this restriction would be comparable to only allowing an examinee to read an item stem one or two times; it would be regarded as unnecessary or even inappropriate.

Decisions about the level of examinee control over the audio that will be provided should be based on the construct area. These decisions will directly impact how the item type will be implemented in the CBT software, but they will also have far reaching implications for characteristics of how the audio, and related items, should be written. Item writing guidelines should be specified that support the measurement goals.

Realism

Another area where decisions may need to be made concerns the "realism" of the audio files. Realism in audio will often conflict with clarity and simplicity. For example, in non-speech files, ambient or background noise is typically avoided in order to ensure that the critical element is fully audible. However, in some instances the assessment goal for the construct could turn on whether the examinee is able to identify the critical element under realistic, "noisy" conditions. For speech sounds, additional decisions about realism may address the extent to which regional or international accents are used and whether emotional tone is included.

The potential concerns related to realism should be analyzed for a given exam program and the decisions made should be appropriate for the construct. Once these decisions have been made they can be incorporated into item writing guidelines that the item writers follow as they produce each new test item.

The Hot Spot Item Type

The issues for hot spot items are very different from those that impact audio items, although in both of these example instances the concerns largely stem from the use of media. Most of the issues in audio items relate to the use of sound, while most of the issues for hot spot items concern their use of images.

Technical Characteristics of Images

The hot spot item type is only one of several item types that include images. Traditional items, such as the multiple choice and the multiple response, can be written to display graphics in either the stem or the response options. An exam program may elect to impose some general decisions regarding the use of images in any item type. These decisions might include technical constraints regarding the file type or image size permitted. They might also apply to specific characteristics of the appearance of the images, set as the standard for that exam.

Attributes of Correct and Incorrect Graphical Areas

In addition to any general decisions about the use of images, hot spot items have a few specific considerations, due to the fact that in this item type the graphic itself serves as the response interface for the item. The image in a hot spot item must include a key area within the graphic. Incorrect responses to a hot spot item may be presented as specific graphical distractor areas; alternatively, the item could include a correct area while the remainder of the image could be treated as a general undefined incorrect graphical area.

Since the image itself includes the key and distractor elements, additional guidelines may be warranted. An exam program may establish a guideline that in every hot spot graphic the correct area must be highly distinct from the incorrect areas, as well as from the remainder of the image. This reasonable guideline might have implications for both the appearance and the scale of the image. For example, a hot spot item might be written in which the key is the nation of Belgium. If the graphic used is a political map of Europe, then each country's boundary lines can serve as clearly distinct and discernable response options. On the other hand, if the image used is instead a relief map, it might be very difficult for an examinee to be sure he or she has selected the correct element. Furthermore, if the relief map were of the hemisphere, rather than just Europe, the size of the correct area might be too small for an examinee to fully control mouse selection, even if he or she knew the correct answer.

Response Markers

There are two primary ways in which hot spot items may function within a CBT application. Any item writing guidelines determined for an exam program will need to be based on which hot spot functionality is available in the CBT software in use.

In one approach to hot spot functionality, specific aspects of the hot spot image become distinct when an examinee moves the mouse over the image. For example, an image of a piano keyboard might be coded so that each piano key appears individually highlighted when the user moves the mouse over that part of the image. Each of these highlighted areas indicates a distinct response option within the graphic. Each of these graphical response areas must be defined with the software by a test developer, in advance of testing. If an exam program is being delivered on CBT software with this functionally, a reasonable guideline might be that each hot spot image should include at least four distinct areas which the item writer specifies as the key and plausible distractors.

An alternative functioning of the hot spot item type, implemented in other CBT applications, does not display any highlighting. In this approach, the hot spot image does not change as the examinee moves the mouse over the image. The examinee makes a selection by clicking on an area of the image. The CBT software then displays some clear symbol or marker on top of the hot spot image, to mark the place where the examinee clicked. For this type of hot spot functionality, decisions might be made within the exam program about issues related to the examinees' interactions with the image. For example, the size of the key area in each image could be required to be larger than the marker itself. Otherwise, when an examinee selects the correct

area the "response marker" may extend into other, incorrect, areas of the image. This could cause some examinees to worry about how their response will be read and scored by the CBT software.

All of these decisions about graphics and hot spot items will potentially impact the types of items that may be written for an exam program as well as the details of each item's implementation.

Training of Item Writers

Item writer training is a key component in ensuring the development of high quality items (Downing, 2006; Schmeiser & Welch, 2006). As with item writing guidelines, formalized item writer training is provided in the great majority of traditional paper-and-pencil exam programs. However, due to the many new elements in innovative item types, new or expanded item writer training is likely to be needed. Even experienced item writers will need to be given further instruction and practice opportunities when writing new items types.

The revised item templates and the expanded item writing guidelines should be incorporated into the new item writer training materials. The type of innovation employed, as well as the number of new guidelines and the characteristics of the item templates, will all impact the extent to which item writer training will need to be expanded. Expansion of the item writer training materials may include instructions related to the use of templates, examples of the new types of items, and in some cases, modifications to the item writing procedures.

The use of prototype innovative items can be extremely helpful in item writer training. Without a prototype, it can be difficult for item writers (particularly those who are experienced at writing text-based items) to envision what the completed new item might look like and how it might function. A simple prototype, even if it is developed in a rudimentary fashion, can be extremely beneficial; a fully developed prototype will help even more. During the training session, providing item writers with examples of prototypes, along with samples of completed templates associated with each of those prototypes, will allow them to see the full process. For example, when creating a hot spot item, the item template may specify the examinee instructions, location of the image area on the screen, and the minimum and maximum number of selectable regions within an image. This template could be further defined in a sub-template to specify that a particular minimum set of areas within the image be selectable in each item that is created from the template. Useful prototype items built from this template might consist of different appropriate images, or modifications to the areas within an image that can be selected.

The objective in creating samples of completed item prototypes should be to illustrate the range of variations that item writers can make when completing items based on a template. The prototypes and related training should make clear the elements that item writers can change (such as the basic image), and those that will remain constant across all instantiations of the item type or template (such as instructions, selection indicator, feedback, etc.).

These sample item prototypes may be especially important when training item writers tasked with creating audio or video items. Showing a completed template (i.e., script and production specifications), along with the audio that was produced based on the script, can save an immense amount of time explaining what item writers need to create. Having item writers create scripts and practice speaking them with the appropriate inflection, timing, and tone can also be a helpful training exercise when audio items are needed.

The complete design of a new item type is an iterative process. For this reason, existing item development processes may also need to be refined iteratively. Revising and expanding item writer training may require several stages of testing and refinement. However, time spent developing clear procedures and expanded training should result in increased efficiency of item production for these new item types, as well as better quality.

Usability Testing

In addition to the development of the three types of item writing materials described above, the task of usability testing can substantially improve the quality and effectiveness of new item types. Usability testing, unlike the other three tasks presented in this paper, does not have a lengthy history in assessment. Nevertheless, it has a well established reputation in the field of Human-Computer Interaction (HCI) design (e.g., Kirakowski & Corbett, 1990; Shneiderman & Plaisant, 2005). The general principles of usability testing are well known and its effectiveness in improving software has been repeatedly proven and documented (e.g., Bias & Mayhew, 1994; Dumas & Redish, 1999; Gould, Bois, & Ukelson, 1997; Karat, 1997; Landauer, 1995; Nielsen, 2003; Tullis, 1997). Furthermore, one primary approach to usability testing, the "think aloud" method, is based on methods for assessing cognitive processes (e.g., Ericsson and Simon, 1993). And the purposes and goals of usability testing have much in common with Universal Design methods (e.g., Harms, Burling, Way, Hanna, & Dolon, 2006; Johnstone, Thompson, Bottsford-Miller, & Thurlow, 2008) in their emphasis on improving the accessibility of a software interface.

In the following paragraphs some of these basic aspects of usability studies are detailed, along with illustrations of their application in measurement settings. (An illustration of an informal usability test, in which the user attempts a dozen computer tasks using a Linux operation system, is available online at http://contentconsumer.wordpress.com/).

Basics of Usability

Usability is an important aspect of any software application, as it is the program's relative easiness to learn and to use. *Usability testing* is a kind of small sample research study designed to evaluate a software program in terms of potential usability problems (Nielsen, 2003). Testing for usability has become a routine procedure in many companies as part of the development process for new or revised software applications. Usability studies have consistently been proven to improve software so that it can be learned more quickly, is more efficient to use, and results in fewer user errors (Landauer, 1995). In addition, when a usability approach is incorporated into the software development process early on it produces financial benefits, as usability testing reduces programming, prioritizes development, and reduces maintenance and support (Karat, 1997).

Usability studies typically focus of those aspects of a software application known as the user interface. The user interface comprises the elements of a software program that the user sees and interacts with – in other words, the functions available to the user, the forms of navigation, the level and types of interactivity, the visual style, the screen layout, and the written communication to the user. Even minor improvements in the usability of these elements can produce important effects such as a substantial reduction in software use errors (Tullis, 1997). Examples of worthwhile guidelines and principles for software usability can be found in Dumas and Redish (1999) and in Shneiderman and Plaisant (2005).

Good usability is critical for computer based tests because poor usability can be a source of measurement error (Harmes & Parshall, 2000). CBTs which include innovative items have an even greater need for good usability, since innovative item types often present examinees with more complex tasks and interactions than those they experience with multiple-choice items (Parshall, Spray, Kalohn, & Davey, 2002).

The importance of CBT usability has been addressed by a number of measurement professionals (Bennett & Bejar, 1998; Bunderson, et al., 1989; Millman & Green, 1989; Parshall & Harmes, 2005) through an emphasis on the design of the item screens and the quality of the user interface. Usability testing has been conducted on CBT interfaces in several reported studies (Harmes et al., 2004; Hoffman, Harmes, & Erb, 2007; Kayser & Parshall, 2008; Wendt, Harmes, Wise, & Jones, 2008).

Think Aloud Method

A wide range of usability methods exist, including some that require specialized equipment and facilities (e.g., usability labs). However, effective "discount usability methods" (Nielsen, 2007) are also available. The "think aloud" method is one simple but highly effective approach to usability testing.

In the think aloud usability method, the usability participant is asked to speak out loud as he or she attempts to use the software to carry out realistic tasks. The participant's comments are noted and his or her software interactions are observed. A typical study requires at least a two-person team in which one person manages the session and guides the participant, while the second person records important comments made as the participant thinks aloud. Both members of the team also observe the participant's subjective reactions to the software throughout the usability test. In particular, the administrators watch the participant for any signs of confusion or frustration, any uncertainty as to where to look on the screen, or any expressions of satisfaction or success. The overall process is very useful in revealing areas of difficulty or confusion that users may experience with the software.

Think aloud usability studies have been conducted on various aspects of item or test design (Harmes et al., 2004; Hoffman, Harmes, & Erb, 2007; Kayser & Parshall, 2008). For example, in Harmes et al 2004, (as described in Harmes & Parshall 2007) think alouds provided insight into key measurement aspects of the innovative item type design. In Hoffman, Harmes, & Erb (2007), think alouds were used to examine the software used to develop items and test specifications. In Kayser & Parshall (2008) think alouds were used to investigate the effectiveness of instructional screens in preparing examinees to use several innovative item types.

The think aloud method can also be used to identify aspects of participants' cognitive processing. This type of study could provide data that would help in the design of an innovative item type as well as potentially contributing to its validity evidence. A study by Wendt, Kenny, & Marks (2007) used the think aloud protocol to investigate whether certain novel item formats tapped higher order thinking to a greater extent than multiple choice versions of the same items.

Early Prototyping

Whenever possible, it is highly beneficial to begin usability testing very early in the design process. Early usability testing can inform the design decisions before extensive development has already occurred, reducing the number of programming changes needed. Furthermore, early identification of usability problems can enable test developers to improve each item type design quickly and cost-effectively (Bias & Mayhew, 1994).

In order to conduct this early usability testing, it is often valuable to develop low-fidelity "mock-ups" or prototypes of the software screens and functions under consideration. In fact, in many cases prototypes provide the first realistic opportunity for development staff to fully envision what the implementation of a new item type might entail. As such, they often have great value in furthering design decisions.

For CBT innovative items, prototype item screens can be developed fairly easily in software applications such as *PowerPoint* (e.g. Kayser & Parshall, 2008). These prototypes do not need to be fully functional; however, they should reflect any item type features or characteristics that warrant usability testing. Even paper-and-pencil prototypes, along with textual descriptions of specific functionality or implementation details, can be very useful at an initial stage. At the other extreme, when a custom software application is being used, testing early item prototypes can serve as a test of the item-level as well as the test-level functionality (Harmes et al, 2004; Wendt, Harmes, Wise, & Jones, 2008).

Multiple Rounds

A consistent principle of usability studies is that multiple rounds of usability testing ought to be conducted (Gould, Bois, & Ukelson, 1997). Multiple rounds are recommended for several reasons. First, this enables usability testing to begin early in the design process, when initial decisions need to be informed by quickly obtained data, while also providing for the review of later design decisions. Furthermore, follow-up rounds of usability tests allow study designers to investigate potential solutions to problems that were revealed in earlier rounds. In other words, if a usability problem is identified in one round, then a revised screen design or other possible solution can be implemented and evaluated in the next usability round. Finally, certain usability problems will not be uncovered until other problems have been resolved. Thus, follow-up studies, as the overall design is being improved, are able to reveal deeper usability problems (Nielsen, 2000).

CBT applications of usability testing with multiple rounds (Harmes & Parshall, 2000; Harmes et al., 2004; Kayser & Parshall, 2008; Wendt et al, 2008) have also documented the effectiveness of this usability design principle.

Design of the Usability Study

The goals and concerns for a given software application provide the foundational basis for the design of the usability tests (Dumas & Redish, 1999). Each round of usability testing should be structured so that the participants undertake targeted software use tasks. These tasks should be selected and designed so that specific goals and concerns related to the software application can be investigated. It is necessary to structure the design of usability tests in part because of time constraints. A think aloud protocol can be time consuming and it is unreasonable to expect a participant to stay focused and effective for more than a couple of hours. The design of the usability test will help ensure that the most critical aspects of the software interface are evaluated.

For a CBT, usability concerns might include the examinees' understanding of navigation functions, the readability of any written instructions, and the clarity of each item interface. The design of a usability study to target these goals might address the item interfaces for any new or innovative item types in the first round, while the follow-up rounds could build on the initial findings and add additional tasks related to test navigation, use of tutorials or help screens, or other specific software concerns (Kayser & Parshall, 2008). Alternatively, each round could investigate additional instantiations of innovative tasks (Harmes et al., 2004; Wendt, Harmes, Wise, & Jones, 2008).

Number of Participants

The optimal number of participants for a single round of usability testing is surprising low. With only five participants, approximately 85% of the usability problems in a software application can be identified (Nielsen, 2000, 2006). In general while the first few participants are highly informative, providing a great deal of unique information, later participants tend to identify the same usability results already noted. As the number of participants increases, the overlap between the usability findings revealed by each becomes more evident. The recommended approach is thus to limit the number of participants in each round of usability testing, while devoting resources instead to conducting multiple rounds.

Characteristics of Participants

The selection of participants for a usability study is a critical concern. According to Dumas and Redish (1999), "One of the cardinal rules of usability testing is that the people who work with the product in the usability test must be like the people who will actually use the product." For CBTs, this means that the usability study participants should either come directly from the examinee population or be highly similar to the expected test-takers.

Once a general category of usability study participants has been identified, further relevant characteristics may also be considered. For any CBT application, one additional characteristic that may be relevant is the participants' level of computer experience. Other examples of participant characteristic that might be important include language background, reading skills, test anxiety, gender, or ethnicity. If a characteristic is deemed to be relevant, then the usability study participants should include some individuals with that characteristic.

In Kayser & Parshall (2008), the critical characteristics of the participants included computer skills, reading ability, and native language. Thus, usability study participants were obtained so that each round included some individuals who had little to no computer experience, who had low reading ability, or who were non-native speakers of English. Targeting the participants in this manner revealed specific problems which could then be addressed, and gave increased confidence to the generalizability of the findings.

Recording the Usability Study

Most usability studies include some means of recording the event for later analysis. In the past this has often meant that an additional staff person would videotape the computer screen while the participant used the software. An alternative approach that is now available is to use screen-capturing software (e.g., *Camtasia*, www.techsmith.com or *iShowU*, www.shinywhitebox.com). This type of software can record the screen throughout an entire usability test, along with audio of the participant's comments as he or she interacts with the software. The resulting digital video file can re-play the participant's audio comments while displaying the onscreen mouse movements and button clicks, enabling a more detailed examination of the participant's interactions or comments. Concurrent use of an external audio recording device, such as a digital voice recorder, can also be used. This second audio recording provides a backup for the audio file saved by the screen-capturing software and can help ensure that the participant's voice comes through clearly enough to be accurately heard or transcribed. This follow-up review phase, using both screen captured video and related audio comments, can be very helpful as decisions about software revisions are being made.

Usability tests conducted on CBTs have recorded the studies through differing means. Harmes et al, (2004) used videotaping, while Kayser and Parshall (2008) and Hoffman, Harmes, and Erb (2007) used screen-capturing through the *Windows*-based software program *Camtasia*. Wendt, Harmes, Wise, and Jones (2008) used an *Apple*-based screen-capturing program, *iShowU*, in their usability testing of innovative nursing items. They also used an external digital voice recorder for audio backup. This proved especially helpful with participants with softer voices.

Sample Applications of Usability in Innovative Item Development

Innovative item applications of usability testing have found a range of beneficial results. For example, in the Kayser and Parshall (2008) study, usability testing of both a hot spot item and an audio item led to changes in the instructions for each item type. These revisions produced substantial improvements in the usability of both item types. Similarly, in the Wendt, Harmes, Wise, and Jones (2008) study, usability testing of a prototype hot spot item resulted in changes to visually clarify the boundaries around areas that examinees selected. In the same study, usability testing of an audio item prototype indicated that the item screen needed to be modified to allow examinees to view client information and play audio files simultaneously.

Summary

The measurement field as a whole has acquired considerable expertise in how examinees will interpret and interact with traditional item types. The multiple choice item, for example, was developed and refined over many years of use in a wide variety of exam programs. That deep, broad level of knowledge and understanding is not yet present for innovative items. Research is needed to increase our knowledge of the

psychometric functioning of various item types and our understanding of the best ways to utilize each of them.

A variety of investigations into how these novel item types differ from their text-based counterparts is warranted. Examples of relevant studies that have been conducted include an investigation of measurement efficiency (Jodoin, 2003), research into cognitive processing (Wendt, Kenney, & Marks 2007; Wendt & Harmes, in press), and a consideration of test security implications (e.g., Harmes, Kaliski, & Barry 2007). Other studies of innovative item types have included a feasibility analysis (Zenisky & Sireci, 2001), statistical analyses (Parshall & Becker, 2008), analysis of construct validity (Sireci & Zenisky, 2006), and a cost/benefit comparison (Harmes & Parshall, 2007; Parshall & Harmes, 2008b).

Further research into the range of new item types, their functionality, and their relative contributions towards validity, will continue to expand our proficiency in their design and use. Furthermore, as operational exam programs report more fully on the results of innovative item types, across examinee groups and over time, the characteristics of each item type will become increasingly evident. As that information is disseminated through the measurement field, best practice in terms of each item type's design will become known.

In the meantime, while innovative items are still new and incompletely understood, we believe that the tasks presented in this paper offer helpful guidance for both the design and development of high quality innovative items. For example, when templates and prototypes are used in the planning stages of innovative item type design a thorough internal review process can reveal and address many potential problems at a very early stage. In addition, think aloud studies conducted during item design can uncover usability problems in the software interface. At a somewhat later design stage, other think aloud studies can determine the examinees' cognitive processing as they interact with the new item types. Furthermore, if the expanded item writing guidelines and item writer training are pilot tested on SME committees, and then revised based on feedback collected, substantial improvements can be obtained in the quality of the items written to the new item types. All of these tasks can be conducted in the design stage in an iterative, research-oriented fashion, as a means of learning more about the optimal item type design.

Finally, the activities of template design, item writing guidelines, item writer training, and usability testing are also worthwhile during the development of innovative item types. Templates can improve the structure, efficiency, and security of item writing efforts. Expanding the item writing guidelines and item writer training for any new item types provides definite opportunities for item quality improvements. Finally, usability methods can help ensure that the examinees fully understand how to use the exam interface, as well as any interface elements specific to an item type. When these four tasks are fully incorporated in the test development process then the great promises of improved measurement through innovative item types can be met.

References

Ballas, J. (1994). Delivery of information through sound. In G. Kramer (Ed.), *Auditory Display* (pp. 79-94). Reading, MA: Addison-Wesley.

Bennett, R. E., & Bejar, I. I. (1998). Validity and automated scoring: It's not only the scoring. *Educational Measurement: Issues and Practice, 17*, 9-17.

Bias, R. G., & Mayhew, D. J. (Eds.). (1994). *Cost-justifying usability*. Boston: Academic Press.

Bunderson, V. C., Inouye, D. I., & Olsen, J. B. (1989). The four generations of computerized educational measurement. In R. Linn (Ed.) *Educational Measurement*. 3rd edition. New York: American Council on Education and Macmillan Publishing Co.

Content Consumer. (2008). *The great Ubuntu-girlfriend experiment.* Retrieved April 28, 2008 from http://contentconsumer.wordpress.com/2008/04/27/is-ubuntu-useable-enough-for-my-girlfriend/

Downing, S. M. (2006). Twelve steps for effective test development. In S. M. Downing & T. M. Haladyna (Eds.). *Handbook of Test Development* (pp. 3-25). Mahwah, NJ: Lawrence Erlbaum Associates.

Dumas, J. S., & Redish, J. C. (1999). *A practical guide to usability testing, Revised Edition*. Exeter, England: Intellect.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data (Revised edition).* Cambridge, MA: MIT Press.

Gaver, W. W. (1989). The SonicFinder: An interface that uses auditory icons. *Human-Computer Interaction, 4*, 67-94.

Gould, J. D., Bois, F. J., & Ukelson, J. (1997). How to design usable systems. In Helander, M., & Landauer, T.K., & Prabhu, P. (Eds.), *Handbook of human-computer interaction, 2nd, completely revised edition.* (pp. 231-254). New York: Elsevier Science Publishers.

Haladyna, T. M. (1996). *Writing test items to evaluate higher order thinking*. Needham Heights, MA: Allyn & Bacon.

Harmes, J. C., Kaliski, P. K., & Barry, C. L. (2007, November). *Are they really more memorable? Implications of innovative items for test security*. Paper presented at the annual meeting of the Florida Educational Research Association, Tampa, FL.

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

Harmes, J. C., & Parshall, C. G. (2007, February). *Development and evaluation of an innovative computer-based assessment.* Poster presented at the annual meeting of the Association of Test Publishers, Palm Springs, CA.

Harmes, J. C., Parshall, C. G., Rendina-Gobioff, G., Jones, P. K., Githens, M. P., & 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.

Harms, M., Burling, K., Way, W., Hanna, E., & Dolon, R. (2006, April). *Constructing innovative computer-administered tasks and items according to Universal Design: Establishing guidelines for test developers.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Hoffman, D. J., Harmes, J. C., & Erb, J. P. (2007, April). *Usability evaluation for computer-based testing software:* Comparing *method effects on information acquisition.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Jodoin, M. G. (2003). Measurement efficiency of innovative item formats in computer-based testing. *Journal of Educational Measurement*, *40*(1), 1-15.

Johnstone, C. J., Thompson, S. J., Bottsford-Miller, N. A., & Thurlow, M. L. (2008). Universal design and multimethod approaches to item review. *Educational Measurement: Issues and Practice*, 27(1), 25-36.

Karat, C. (1997). Cost-justifying usability engineering in the software life cycle. In Helander, M., & Landauer, T.K., & Prabhu, P. (Eds.). *Handbook of human-computer interaction, 2nd, completely revised edition*. (pp. 231-254). New York: Elsevier Science Publishers.

Kayser, M., & Parshall, C. G. (2008, March). *Building a global innovative test*. Presented at the annual meeting of the Association of Test Publishers, Dallas, TX.

Kirakowski, J. & Corbett, M. (1990). *Effective methodology for the study of HCI.* New York: North-Holland.

Landauer, T. K. (1995). *The trouble with computers: Usefulness, usability, and productivity*. Cambridge, MA: MIT Press.

Millman, J. & Greene, J. (1989). The specification and development of tests of achievement and ability. In Linn, R. (Ed.). *Educational Measurement*. 3rd edition. New York: American Council on Education and Macmillan Publishing Co.

Nielsen, J. (2000). *Why you only need to test with 5 users.* Retrieved November 4, 2004, from: http://www.useit.com/alertbox/20000319.html

Nielsen, J. (2003). *Usability 101: Introduction to usability.* Retrieved April 4, 2008 from http://www.useit.com/alertbox/20030825.html

Nielsen, J. (*2007*). *Fast, cheap, and good: Yes, you can have it all.* Retrieved April 4, 2008 from http://www.useit.com/alertbox/quantitative_testing.html

Nielsen, J. (*2006*). *Quantitative studies: How many users to test.* Retrieved April 25, 2008 from http://www.useit.com/alertbox/fast-methods.html

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

Parshall, C. G. & Becker, K. A. (2008, July). *Beyond the technology: Developing innovative items*. Presented at the bi-annual meeting of the International Test Commission, Manchester, UK.

Parshall, C. G., & Harmes, J. C. (2008). The design of innovative item types: Targeting constructs, selecting innovations, and refining prototypes. *CLEAR Exam Review, 19(2).*

Parshall, C. G. & Harmes, J. C. (2008, March). *Stages in designing innovative item types*. Presented at the annual meeting of the Association of Test Publishers, Dallas, TX.

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

Parshall, C. G., Harmes, J. C., Davey, T., & Pashley, P. (In press). Innovative items for computerized testing. In W. J. van der Linden & C. A. W. Glas (Eds.). *Computerized adaptive testing: Theory and practice, 2nd Edition*, Norwell, MA: Kluwer Academic Publishers.

Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). *Practical considerations in computer-based testing*. New York: Springer-Verlag.

Roid, G. H., & Haladyna, T. M. (1982). *A technology of test-item writing.* New York: Academic Press.

Schmeiser, C. B. & Welch, C. J. (2006). Test development. In R. Brennan (Ed.). *Educational Measurement 4th edition*, (pp. 307-353). Westport, CT: Praeger Publishers.

Shneiderman, B., & Plaisant, C. (2005). *Designing the user interface: Strategies for effective human-computer interaction*. Boston: Pearson/Addison Wesley.

Sireci, S. G., & Zenisky, A. L. (2006). Innovative item formats in computer-based testing: In pursuit of improved construct representations. In S. M. Downing & T. M. Haladyna, (Eds.), *Handbook of Test Development* (pp. 329-347). Mahwah, NJ: Lawrence Earlbaum Associates.

Tullis, T. (1997). Screen design. In Helander, M., & Landauer, T.K., & Prabhu, P. (Eds.), *Handbook of human-computer interaction, 2nd, completely revised edition.* (pp. 503-531). New York: Elsevier Science Publishers.

Wendt, A., & Harmes, J. C. (in press). Developing and evaluating innovative items: Part II, item characteristics and cognitive processing. *Nurse Educator*.

Wendt, A., Harmes, J. C., Wise, S. L., & Jones, A. T. (2008, March). *Development and evaluation of innovative test items for a computerized nursing licensure exam.* Paper presented at the annual meeting of the American Educational Research Association, New York.

Wendt, A., Kenny, L. E., & Marks, C. (2007). Assessing critical thinking using a talk-aloud protocol. *CLEAR Exam Review*, *18*(1), 18-27.

Zenisky, A. L., & Sireci, S. G. (2001). *Feasibility review of selected performance assessment item types of the Computerized Uniform CPA Exam.* (AICPA Research Consortium- Examinations Team. Technical Report) AICPA: Author.