Beyond the Technology: Developing Innovative Items

Cynthia G. Parshall, Ph.D
Measurement Consultant

Kirk A. Becker, Ph.D
Pearson VUE

Abstract

In the last decade, it has become technically feasible for testing organizations to provide innovative items for computer-based assessment as a standard offering. Several item types, including hotspot, drag-and-drop, and fill-in-the blank have been available for several years, but are not in widespread use. Our experience with these item types has shown that implementing the following processes improves the quality of the assessment and provides a framework for developing these items as part of the normal test development process:

- Develop specific materials (i.e., item templates, item writing guidelines, and training for item writers) for each innovative item type.
- Conduct several rounds of usability tests with a representative sample of candidates early in the test development cycle to ensure that examinees will understand each item type.
- Analyze both correct and incorrect responses from pretest data for each innovative item type.

In this paper we discuss each of these recommended practices with illustrative examples from exam programs such as the Driver Certificate of Professional Competence (CPC), the Inventory of Teacher Technology Skills (ITTS), and the Pearson Test of English (PTE).

Introduction

This paper looks at what steps, beyond technology, need to be in place for innovative assessments to be valid and reliable. Our objective is to delineate best practices for developing innovative item types. While innovative items appear to hold great promise for improving our overall measurement (Parshall, Harmes, Davey & Pashley, in press), it is also evident that thoughtful, deliberate design is needed to fully realize this potential.

The additional effort for innovative items is needed because of a lack of fundamental research on how test takers will interpret and interact with these new item types. Familiar item types like the essay and the multiple choice have been developed and refined through their use in many exam programs over many years. The measurement quality of these traditional assessment methods can be almost guaranteed through careful adherence to well-known item writing guidelines and test development principles. A new item type, on the other hand, will need time and attention to develop the best approach to its design and implementation.

The goal for an exam program might be to have the item writing process for a variety of item types so well defined that a new item can be easily written using the item type that is most suitable, based on the appropriate match of the content of that new item to the item type.

A Model for Designing Innovative Item Types

One thorough approach to innovative item type design involves a 6-step process that includes several rounds of review and revision (Parshall & Harmes, 2008a). The steps in this approach 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 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. Even content areas that are well addressed by multiple-choice items may benefit from innovative item types. In an ideal setting, the most appropriate item type for a specific purpose can be used for each item. 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. Considering these two steps in this order should yield a better solution than an approach which begins by identifying the types of innovative items available and then deciding how they can be incorporated into the exam program.

In Step 3, *design initial prototypes for internal discussion,* the test developers, in collaboration with Subject Matter Experts (SMEs), begin to define the new item types for the exam program. This item type specification can require substantial cognitive work, depending on the type of innovation and the construct needs. Even relatively simple innovations, such as changing selected response items to free entry, adding graphics to item stems, or utilizing touch screens as the input device can benefit from careful analysis. More elaborate innovations, such as those that include multi-step situated tasks, interactivity, or complexity, warrant a thorough analysis of possible approaches and their implications. Once a preliminary item type design has been specified, then a prototype is mocked-up for preliminary consideration by exam program stakeholders.
These stakeholders will include psychometricians who may consider scoring implications, software development staff who may address the feasibility of each prototype, and usability personnel who may review the prototypes for their usability. 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,* involves three sets of activities. To begin the step, item writing materials will usually need to be developed for each proposed innovative item type. These item writing materials will usually include item type templates, expanded item writing guidelines, and training (Parshall & Harmes, 2008b). A few SMEs can utilize these newly developed item writing materials as they produce a few content-relevant sample items. These sample items should address all the basic features of each planned new item type, so that all item design elements can be considered. The sample items can be examined in detail by exam program stakeholders to further consider each proposed item type's potential. The sample items can also be presented to a few members of the examinee population for *usability testing*, in order to identify any usability problems. The results of the usability tests are used along with the stakeholder reviews to suggest refinements and improvements in the item type designs. Several rounds of feedback and review are recommended. Once guidelines have been developed for a given item type, those materials will generally be appropriate and available for future item development with other programs.

In Step 5, *conduct a pilot study*, trialling of the new item types is conducted. The trialling effort 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. This undertaking provides an opportunity to evaluate the full system implementation of each new item type. An important aspect of this step is the item analyses that are conducted on each item and each item type. Distractor analysis can be particularly relevant for investigating the cognitive functioning of novel item types.

In Step 6, *produce final materials*, all of the materials needed to implement each approved item type as part of the regular exam program are fully prepared. All of the documentation regarding each new item type should reflect the final decisions that were made by the various groups. All of the processes and procedures for exam program activities such as item writing, test development, and pool maintenance should be capable of handling any novel elements associated with the new item types.

In this paper we illustrate the application of Steps 4 and 5. The development of item writing materials is addressed briefly, while testing for usability and conducting item analyses for each innovative item type are each discussed and illustrated in more detail.

Item Writing Materials

In order to help item writers with the task of writing innovative items, additional materials should be developed. These materials, including item templates, item writing guidelines, and item writer training, will give SMEs greater facility with each innovative item type. Improving the SMEs understanding of the item types will ease their task and increase the quality of the items they produce.

Templates

Because of the increased complexity in building innovative items (especially those with media), templates are needed, if only for the technical aspects of an item. Any time innovative item types are developed, beyond "boutique" or "show" item generation, templates become critical. When several hundred items, or more, must be generated, then the need to define graphics, sound, video, content, and other aspects of these items becomes logistically challenging. Without templates, the risk of incomplete item information is quite high, requiring multiple requests from item writers and fewer usable items. Furthermore, when the template also addresses content, variations on a theme, or item taxonomies, it can serve to drive down the well publicized costs of these items.

Templates can provide a framework for item construction that can improve item structure, production efficiency, and exam security. Item structure can be improved with templates by standardizing the way in which each new item format will be constructed and presented. Production efficiency can also be improved, as item writers fill in components of a template instead of creating an entirely new item concept each time. Since templates can be used as a tool for quickly developing different versions of an item, depending upon the level of specificity to which the template has been designed, they can also be useful as a security measure.

A primary feature of most templates is a database form, in which item elements that the item writer needs to supply are listed. Templates can also include visual elements through an emphasis on screen layout. For the CPC Exam, the item writers were presented with a sample case study and sample items. These materials helped make the critical elements of each item type salient to the item writers. While this exam program made extensive use of graphics and sound, the item writers were frequently not in a position to submit images or sound files with their items. Rather, the item writers often submitted a text description of the media requirements for the item. Further template form elements were used by internal test development staff to track the item graphics as they were developed, reviewed, and approved by the item writers, and then linked to the items within the item bank.

Item Writing Guidelines

For each innovative item type, new item writing guidelines should be developed. The standard item writing guidelines, currently in use in many exam programs, have a well-established effectiveness in producing high-quality traditional items. In fact, one obvious reason for the success most standardized exam programs have with multiple choice items is the clear guidelines for this item type which many decades of experience have provided. However, most innovative item types include new elements that

are not fully addressed in the existing item writing guidelines. Typical item writing guidelines such as "make sure the question has only one answer" may become more difficult to comply with as the number of response options moves towards open-ended. On the other hand, innovative items that include graphics may be too variable if item writing guidelines fail to specify characteristics of the images and how they should be used in an item. The provision of thorough item writing guidelines should strengthen the standardization and quality of the items written for each new item type.

Table 1 provides item-total correlations for high-frequency observed responses on a hotspot item with multiple correct options. In this example, the item is scored as a partial credit item. The item-writing guideline in this instance is "make sure that all correct options are identified" rather than "make sure that items have only one correct answer". In this item, option "02" is scored, but responses excluding that option are more frequent and more highly correlated with the total score than those including it. In this instance, either the key or the content should be revised, or the item should be discarded.

Table 1. Response Analysis for Multiple Response Hotspot Item

Response	N	Option Mean	Point Biserial
01	17	0.03	0.05
01*02*03*04 (Key)	48	0.09	0.24
01*03	29	0.05	0.13
01*03*04	**103**	**0.19**	**0.35**
01*46*03*04	32	0.06	0.19

Potential item writing guidelines for Hot Spot items, items using Graphics, and items using Audio are listed below. Some of these guidelines might be appropriately adjusted depending upon specific aspects of an exam program.

Hotspot Items

- The critical elements of the graphic should be clearly evident.
- The correct area in a graphical hotspot item should be large enough for easy mouse selection.
- The graphic should not contain too many irrelevant distractions which may confuse the candidate.

Items Using Graphics

- Graphics should be saved as .GIF or .JPEG. Bitmap files should be avoided as they are very large in size.
- The size of the graphic when used as part of a case should be no bigger than 350 pixels by 200 pixels in size.
- When used as part of an item the graphic should be as small as possible while still being readable for the candidate.
- Overly complex graphics should be avoided as they may make it difficult for examinees to identify the essential information.

Audio Items

- Do not use more than 2 speakers (2 people speaking)
- Do not write a script that would result in an audio recording over 15-20 seconds in length. This generally means the script should be no more than 5
- 6 sentences long. Provide a preliminary sentence, with sufficient context to indicate who is speaking in the audio file. (e.g., "A passenger speaks to the bus driver.").

Item Writer Training

Once the item templates and item writing guidelines have been drafted they should be incorporated into item writer training materials. Depending upon the status of the exam program, these training materials may need to be developed for the first time. In other cases, existing content development and item writer training materials may simply need to be expanded, in order to address the proposed new item types. Specific training activities should be designed for any novel item writing task to help item writers learn how to work with each new item type and to produce consistent, high-quality items.

For the CPC, additional training was necessary to help item writers with the case-based structure of the exam, as well as with the various new item types. In this instance, item writers felt most challenged by the case-based structure of the exam. A sample case-study, consisting of a brief scenario and five related items was developed. Each of the five items in the sample case study demonstrated a different item type. This sample case study was used to present the 'concept' of the exam to the item writers. The item writers were then divided into teams and each team produced a case study, complete with several items. The item writers experimented with writing these case studies both 'top-down' and 'bottom-up'; that is, by producing the case and then writing items, and by writing items first and then writing a contextual case. Practice in writing the innovative item types also occurred.

Usability Studies

The *usability* of a software program is its relative easiness to learn and to use. *Usability testing* is the process of evaluating a software program, to identify any usability problems it may have (Nielsen, 2003). Usability is particularly critical for computer-based test (CBT) applications, because of its potential impact on measurement error (Harmes & Parshall, 2000). Furthermore, CBTs that include innovative items have an even greater need for high usability as innovative items often present examinees with more complex tasks and interactions than those they experience with multiple-choice items (Parshall, Spray, Kalohn, & Davey, 2002). In these instances, the importance of good usability for the CBT interface is especially critical.

One recommendation for usability testing is to begin investigations early in development. In many cases early usability testing is enabled through the use of prototype software.
Prototype software enables developers to quickly determine elements that are, or are not, working in the new item types, potentially leading to quick improvements in the item type designs.

In addition to early usability testing, multiple rounds of testing is also recommended. These multiple iterations provide for a sequence of investigations and refinements. Later rounds of usability tests allow study designers to investigate potential solutions to problems that were revealed in earlier rounds. Furthermore, 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).

Due to the overlap in usability problems revealed by each participant, the optimal number of participants for a single round of usability testing is surprisingly low. It has been offered that, with only five participants, approximately 85% of the usability problems in a software application can be identified (Nielsen, 2000, 2006). The recommended approach is therefore to limit the number of participants in each round of usability testing, while devoting resources instead to conducting multiple rounds.

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 participant characteristics that might be important include language background, reading skills, test anxiety, gender, ethnicity, or other variables. If a characteristic is deemed to be relevant, then the usability study participants should include some individuals with that characteristic.

One simple but highly effective approach to usability testing is the 'think aloud' protocol. In this method, the usability participant is asked to 'think aloud' 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. The participant's subjective reactions to the software throughout the usability test are also noted. In particular, any signs of confusion or frustration on the part of the participant, any uncertainty he or she conveys as to where to look on the screen, or any expressions made of satisfaction or success, are all regarded as important data.

Usability testing for the CPC began early in the development cycle and used prototype software. Three targeted rounds of usability testing were planned for this exam. A few of the usability problems identified, and solutions found, are provided next.

Round 1 Design

The first round of usability testing used a sample case study and an example of each of the five planned item types. Six participants were included in the initial round of usability testing. These participants included one person with low computer skills, two individuals who had low reading abilities, and one non-native speaker of English.

The first round of usability testing was specifically designed to investigate users' understanding of the case-based nature of the exam and their ability to interact with each item type. A sample item screen from Round 1 is displayed in Figure 1.

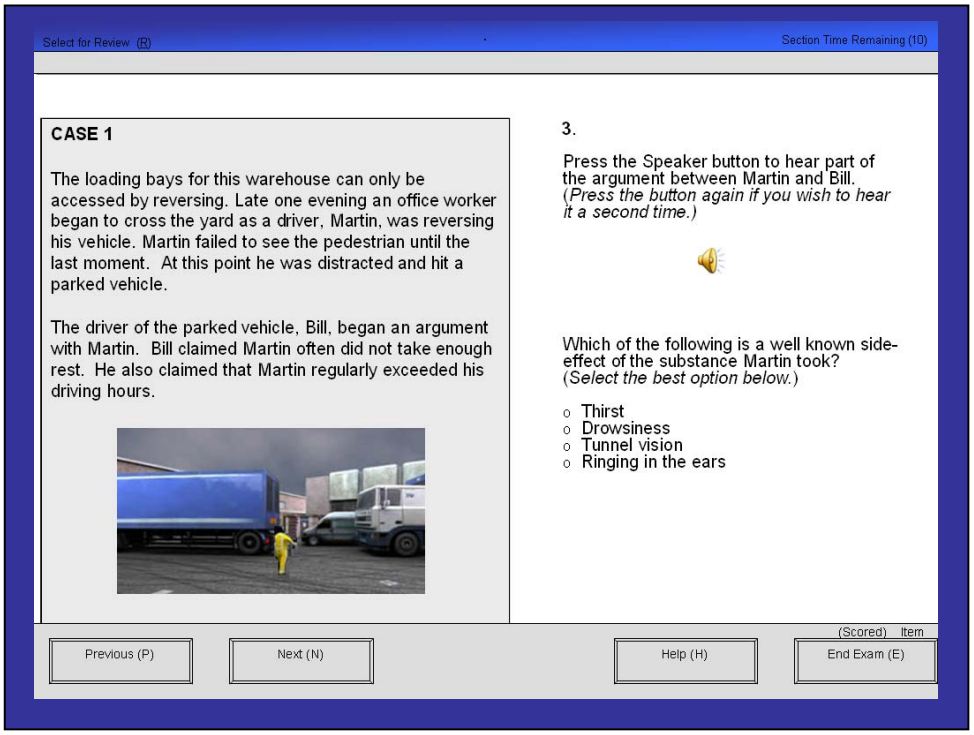

Figure 1. Round1 Sample Item Screen.

Round 1 Findings

Overall reactions from the usability participants at this stage included a number of positive comments.

- Several participants stated that they 'liked the different types of items.'
- Several participants made comments about liking the realistic scenario, and that it was 'good that it was based on the real thing.'
- One participant liked that the scenario 'was always there' and could be referred to again.
- Several participants made comments directly linking the items to the case, recognizing that they were related (e.g., 'It tells you something, and then asks you questions').
- There were positive comments about the use of a photograph in the case. For example, one participant pointed to the photograph and said that 'it really helped you understand' the scenario. (See Figure 1.)
- There were several positive reactions to the use of audio.
- When encouraged to open a Help Screen (no one did so independently) there was universal agreement that the information was clear and helpful.

Overall reactions from the usability participants also included suggestions or concerns. Finding areas that might need improvement is a primary purpose of usability testing so identifying these concerns was a positive outcome.

- One usability study participant and the pilot participant both displayed confusion regarding the first Instruction screen. This screen includes a 'sample screen shot' (reduced in size) of a case and item to clarify the layout of the case structure. (Figure 2 displays this Instruction screen.) However, these two users both thought the sample screen was the actual case -- and they were bothered by the small size, which made it difficult to read the text.
    - A revision of this screen was developed and tested in Round 2. (See Figure 3.)

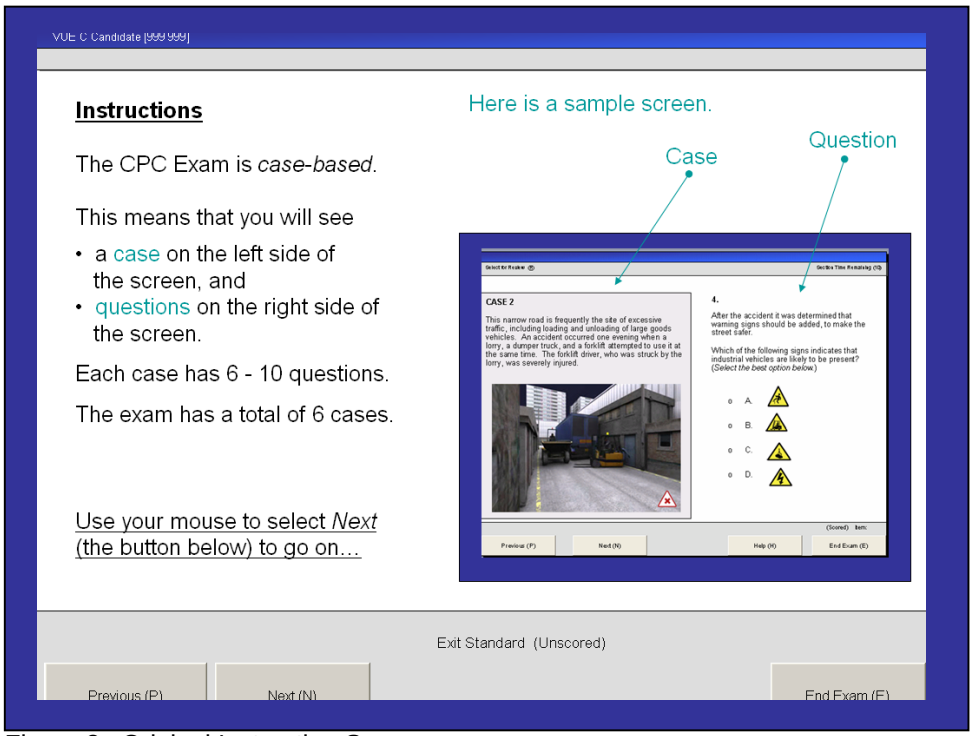

Figure 2. Original Instruction Screen.

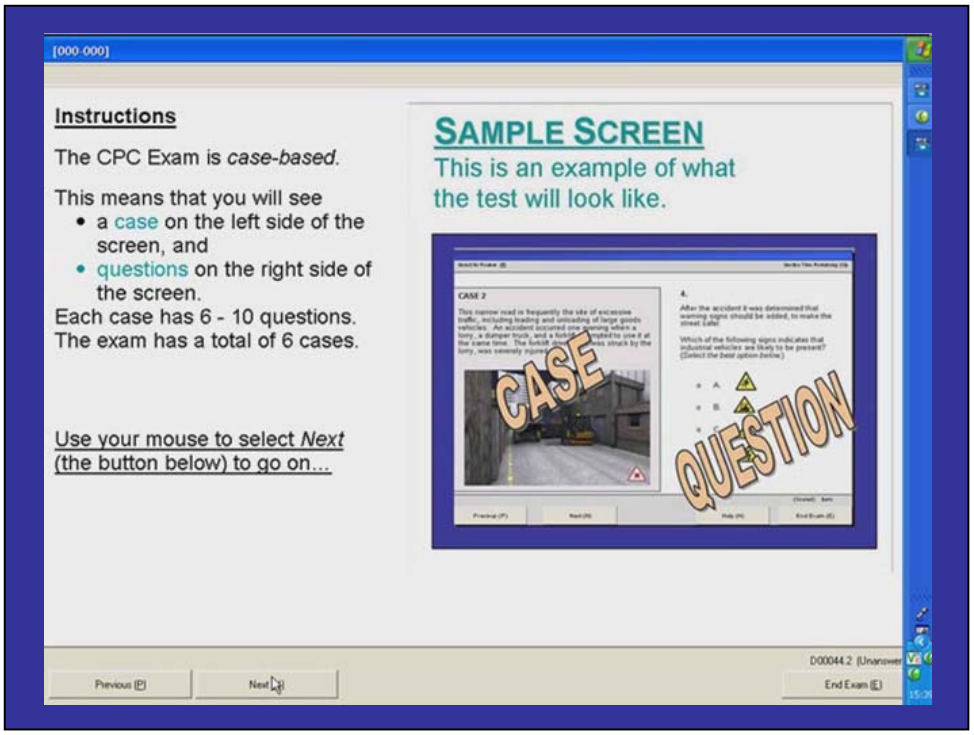

Figure 3. Revised Instruction Screen.

- Several of the examinees only selected one response on the multiple response (MR) item -- even though item instructions stated 'Select TWO …'. (See Figure 4.) Since the prototype software did not force them to select another response, they would have lost credit for these items. For this examinee population, with modest computer skills, it did not appear to help that check boxes were used for MR items and radio buttons for MC items.
    - The embedded instructions were made more prominent on all item screens by making them stand-alone paragraphs. This was investigated in Round 2 and found to be a successful approach. (See Figure 5.)

Round 2 Design

There were a total of 10 participants in Round 2. The participants included one man who referred to himself as 'a dinosaur', to describe his lack of computer experience. Other participants included a native speaker of Somali and a native speaker of Punjabi. In addition, two of the participants demonstrated low reading ability.

Three approved Case Studies were implemented in the actual CBT delivery software for Round, allowing for usability testing of software navigation across cases. A further design goal for Round 2 was to evaluate the effectiveness of the alternative approaches, mentioned above, which were implemented to address usability problems identified during Round 1.

Round 2 Findings

Several positive results were noted in Round 2.

- Participants were able to properly use all MC and MR item types, even those with graphical response options.
- Participants understood the case-based nature of the exam and were able to navigate through the test.

- As in Round 1, no one independently accessed a Help screen, but those participants who were prompted to do so indicated that the information was clear and useful.

The revisions from Round 1 were also both successful.

- The screen shot in the first Instruction screen was made more obviously a 'sample case' screen (see Figure 3 above). With the exception of the Somali-speaking participant, all participants understood both that this was an Instruction screen and the information it is intended to convey.
- The embedded instructions were made more prominent by putting them in stand-alone paragraphs. The MR item type was used properly, with participants providing the correct number of responses. The original version is shown in Figure 4 and the revised version, evaluated in Round 2, is shown in Figure 5.

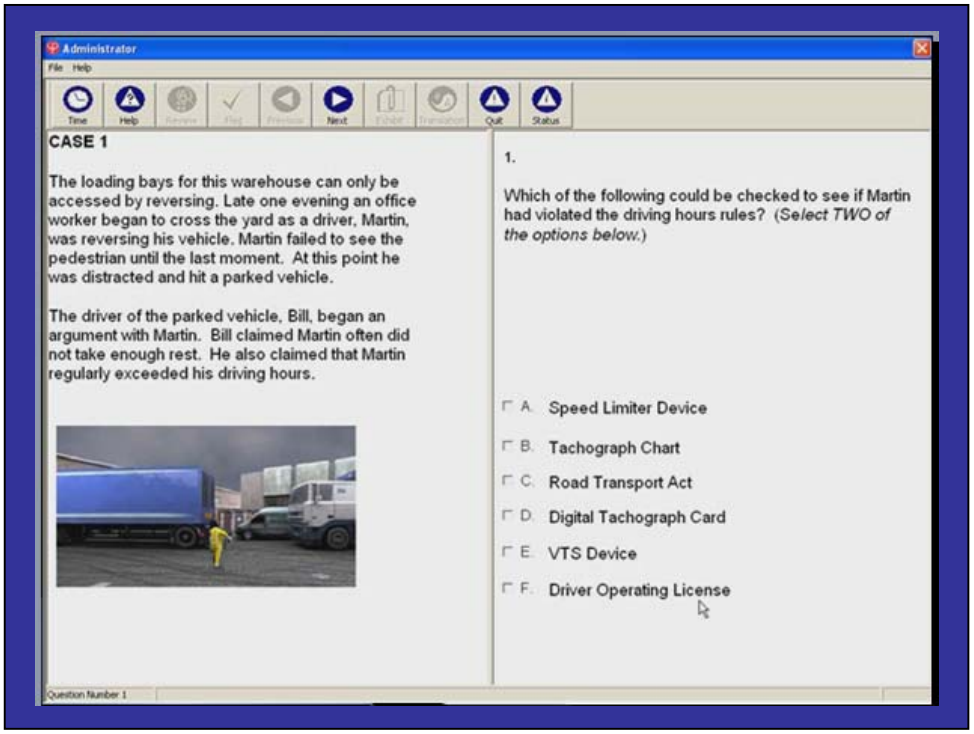

Figure 4. Original Display of MR Embedded Instruction.

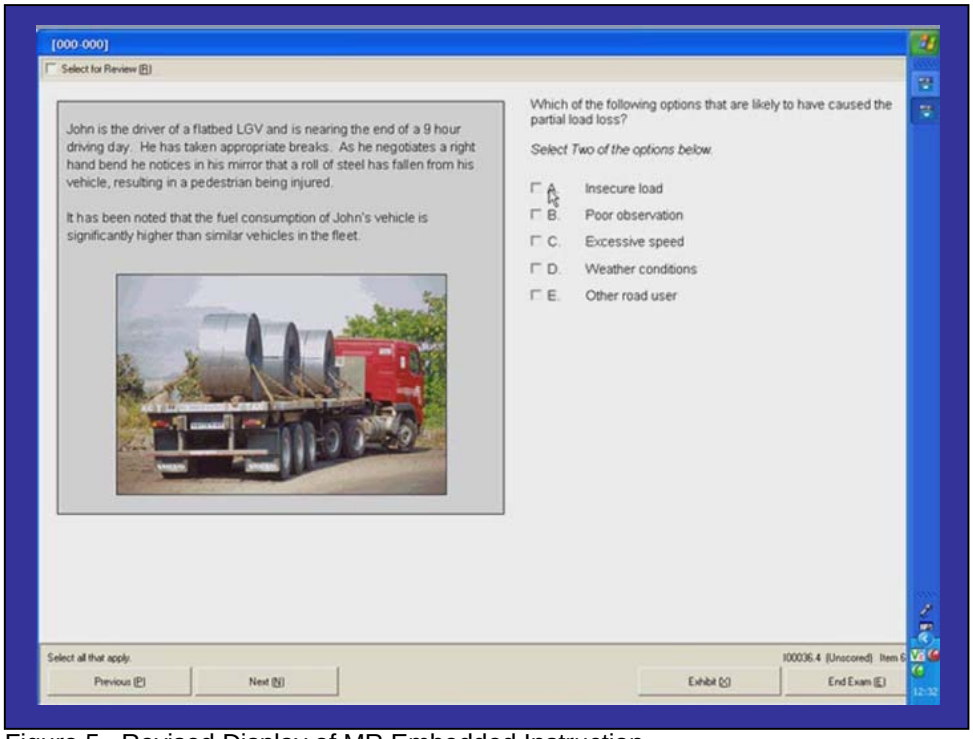

Figure 5. Revised Display of MR Embedded Instruction.

In Round 2 of usability testing for the Driver CPC several distinct usability concerns were identified by the study administrators. Potential solutions for the usability problems were investigated in Round 3.

Round 3 Design

There were a total of seven participants at Round 3, across two days of testing. As a set, the participants spanned the range of skills and abilities that future examinees may be expected to have. One participant stated that his native language was Urdu, although his English skills, both speaking and reading, appeared to be quite good. The participants also included four people who rated themselves as having 'very little' computer experience. One of these participants, after successfully completing the usability test, indicated that he had *never* used a computer before.

A full operational test form, including six approved Case Studies, was implemented in the CBT delivery software for Round 3. The primary goal by this point was simply to determine the effectiveness of alternative design approaches, implemented to address usability problems identified during Round 2.

Round 3 Findings

There were several revisions to the software from Round 2. Most of the changes were minor in nature and are summarized later. One critical change related to an Audio-based innovative item. This concern is addressed next.

- Since several participants in Round 2 had difficulty with the Audio player (see Figure 6), an audio Practice Item was added to the beginning of the test and the embedded instruction was revised. Figure 7 displays this Practice Item, which was evaluated on the first day of usability testing for Round 3. Two of the participants found this audio item to be very clear and usable, but one participant initially failed to see the Audio Player. A further improvement to the Practice Item was

implemented immediately on the second day of Round 3 usability testing (see Figure 8). With this final refinement, all participants were able to successfully use the Audio item.

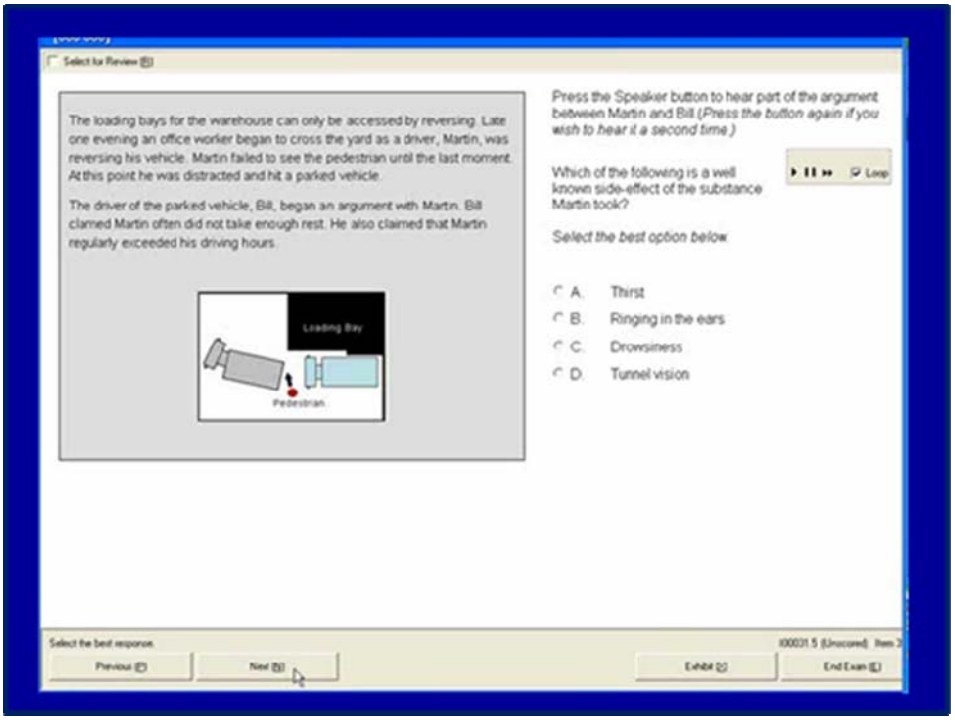

Figure 6. Original Audio Item.

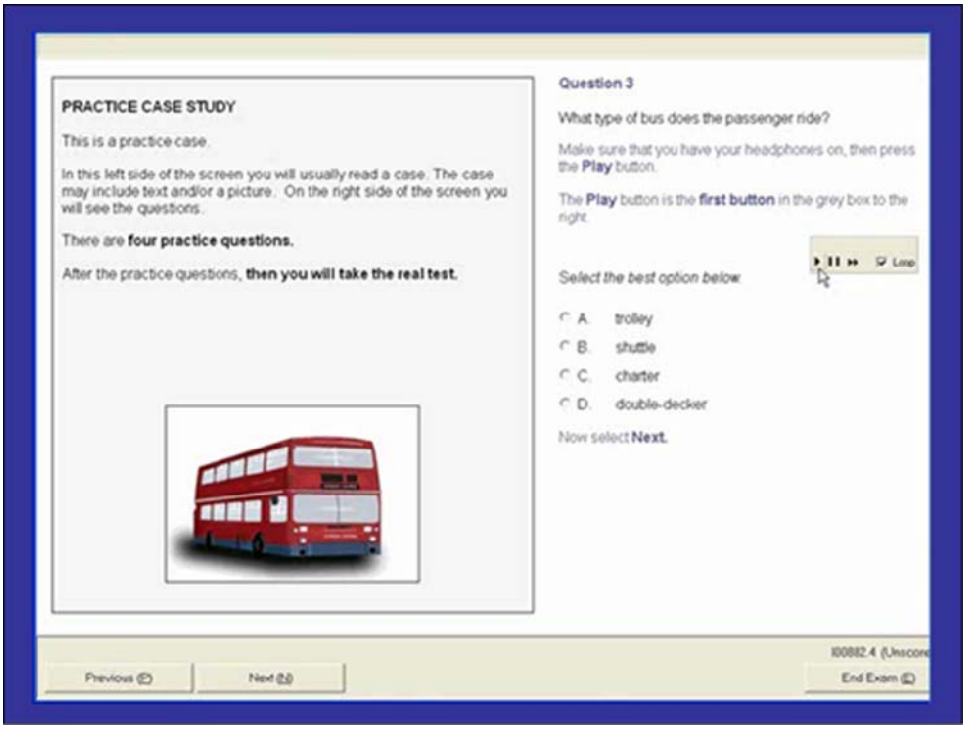

Figure 7. Audio Practice Item.

Figure 8. Revised Audio Practice Item.

The remaining changes from Round 2 were all fully successful as well.

- Participants in Round 2 understood how to use the CBT software to respond to the Hot Spot items, however, several of them were confused once they had responded. This item type displays a red X wherever a user clicks; some examinees interpreted this as a sign of 'getting the question wrong'. In Round 3 a Practice Item for Hot Spots was added to the beginning of the test and the embedded instruction was revised to clarify this issue.
- In Round 2, participants expressed some difficulty with reading the text in the pop-up Message boxes. To address this issue the font size used in these Message boxes was increased from 8 point to 12 point. This greatly improved the visibility of the text in Round 3 and no participant complained or requested a larger font.
- Some participants requested that the buttons be made more prominent. For Round 2, the text labeling each button (e.g., 'Next') was increased to a 12 point font size. This appeared to be fully effective, as no participant in Round 3 had difficulty locating buttons and no one requested more emphasis on the buttons.
- In Round 2 some of the participants indicated a concern that the Case might change from one item screen to the next. To help reassure examinees on this point each Case Enter screen now includes a statement addressing this. In Round 3, no participant expressed this concern; rather, they fully seemed comfortable with the relationship between the Case and the items.
- In Round 2, the two non-native speakers of English both initially failed to read the Case, jumping directly to the questions instead. Two design changes were made in order to help draw the user's eyes to the Case as the best point to begin reading on a given screen. First, a bold header ('Case Study') was placed at the top of each Case. Next, the background on this area of the screen was changed to a paler shade of grey. These changes seemed to address this possible usability concern.

Usability Summary

Usability studies are designed and conducted in order to develop software that can be learned more quickly, is more efficient to use, and results in fewer user errors (Landauer, 1995). The usability testing conducted for the DSA's Driver CPC Exam clearly produced software with improved usability. Three iterative studies were designed, with each study intentionally targeting specific aspects of the CBT software. Several important usability problems were identified, and solutions were designed, implemented, and verified. All of the initial usability goals and concerns were addressed. Furthermore, successful aspects of the CPC Exam were also identified. The CBT software displayed improved usability, with clear instructions and an easy-to-use interface. All of the innovative item types designed for this exam should be accessible to the great majority of exam program candidates. As one participant in Round 3 commented, "It must be pretty easy, because I don't really use computers at all."

Analyses for Innovative Items

In-depth methods for evaluating items, such as examinee interviews and think-aloud protocols, have been used with good success for traditional item types (Pommerich & Burden, 2000; Johnstone, Thompson, Bottsford-Miller, & Thurlow, 2008), as well as innovative item types (Kayser & Parshall, 2008; Wendt, Kenny, & Marks, 2007). Other research methods into specific types of innovative items have included a feasibility analysis (Zenisky & Sireci, 2001), an efficiency analysis (Jodoin, 2003), and a cost/benefit comparison (Parshall & Harmes, 2008b). Until the results from numerous programs are reported it will be difficult to separate the effects of item type from other factors. So long as item writers are uncomfortable with the range of item types, the quality may suffer or item writers may continue to fall back on the standard single-answer multiple choice item type. More general research into item types and their functionality is clearly needed. In this section, this paper will focus primarily on classical and IRT analyses of innovative item types, as well as practical issues related to the evaluation and revision of innovative items over time.

Evaluation of Innovative Item Types

For a program interested in the large-scale implementation of innovative item types, it is critical to understand how different item types perform. Evaluating the comparative quality of innovative item types requires access to accurate classifications of item type in conjunction with the item statistics. A decision may need to be made at the start of the item development process on how item type will be coded – for example, multiple-choice items may contain media stimuli (e.g., audio stems or graphic responses), that are of interest for the evaluation of those items. Especially given the cost of producing media, it is useful to know how these items compare to multiple choice items without a media component. Table 2 provides an example comparison of different item types.

Table 2. Comparison of Item Statistics by Item Type

Item Type	N	Avg Pval	Avg PTME	% Dropped
Multiple Choice	293	80%	0.290	25%
MC w/media stem or response	38	81%	0.293	29%
Multiple Response	114	78%	0.340	13%
Hotspot	30	68%	0.260	20%
Grand Total	475	79%	0.301	22%

Item statistics reported by item type, while useful, should be interpreted with caution. Especially for new programs, item type can be confounded with item authors, content areas, and usability issues. Additionally, decisions on the use of innovative item types in general, or specific innovative item types, may well depend on factors in addition to average performance. Bearing this in mind, the multiple response items in Table XX appear to be functioning better than the other item types.

In addition to tracking costs, performance of different item types can help plan future item development. With information on item mortality by item type, the number of pretest items necessary to achieve the desired number of approved items can be more accurately estimated.

Determination of Measurement Models

Innovative item types typically present several potential measurement opportunities from a test taker's response. While any item can be scored dichotomously, items requiring multiple decisions or responses may warrant polytomous scoring. Dichotomous items require smaller samples for statistical analysis, have well accepted methods for analysis, are easily understood by stakeholders, and can be easily integrated into test content specifications. Polytomous scoring takes advantage of the additional measurement opportunities provided by complex items, the score points cover a range of candidate abilities, and they are typically more discriminating (higher point biserials) than dichotomously scored items.

Item Response Theory models for scoring polytomous items include the Partial-Credit Model (Masters, 1982), the Generalized Partial Credit Model (Muraki, 1992), and the nominal response model (Bock, 1972). All of these models require greater sample sizes than their dichotomous counterparts. Yen and Fitzpatrick (2008) describe the nominal response model as a method for "analyzing items in which it may not be clear a *priori* which answer choices reflect greater ability" (p117). The nominal response model should therefore be considered a tool that informs the development of item scoring, rather than a method for scoring and equating tests.

Derivation of Item Scores

The most basic method for determining item scores is through the content-area expertise of the item writer. In the test development process, there should be exceedingly few instances in which there are no *a priori* correct or incorrect answer choices. This does not mean that the best way of scoring the item is

known, nor that all of the correct or incorrect answers are known *a priori*. The relationship between a response or the probability of a response and ability (based on an interim score for the test items) will provide a basis for evaluating the scored and unscored responses.

An example from the Pearson Test of English pilot study will illustrate this point. One item under consideration involved selecting words and phrases from a passage that referred to a particular person or concept (e.g., references to “Jane Goodall”). Item writers identified the words and phrases that they understood to refer to the person or concept (the keys). There was no limit on the number of words or phrases that a person could select. While the final scoring rules for this item were not known initially, the keys allowed us to associate points to each test taker, which were then used to evaluate and refine scoring models in an iterative process.

As the response space of an item increases from four or five on a single-response multiple choice item to dozens or hundreds, the complexity of correctly identifying correct and incorrect responses also increases. There are several analyses that can help to evaluate the appropriateness of the item scoring rules.

Table 3 shows a distractor analysis from Winsteps (Linacre, 2006) for a multiple-answer multiple-choice, a hotspot, and a single-answer multiple choice item. All items have been scored dichotomously, and the table provides the observed responses, score, frequency (number and percent), average IRT theta for people choosing response, standard error, outfit, and point-measure correlation. Results for item 9 show that all but 4 people (the first four rows) correctly selected two responses, and only the scored response “BE” has a strong positive point-measure correlation. Item 28 is a hotspot item and information is not available on incorrect responses. That is, the item writer did not specify distractor ‘regions’ on the hot spot image. While the item is functioning well, it would be better to have some form of usable information on the incorrect selections. Finally item 457, a multiple-choice item, is not functioning well. Option A is scored incorrect, however it has a higher point-measure correlation than the correct option.

Table 3. Distractor analysis from Winsteps for a multiple-answer multiple-choice, a hotspot, and a single-answer multiple choice item

ENTRY NUMBER	DATA CODE	SCORE VALUE	DATA COUNT	DATA %	AVERAGE MEASURE	S.E. MEAN	OUTF MNSQ	PTMEA CORR.	ITEM	
9	A	0	1	1	-2.14		.0	-.29	4	A
	E	0	1	1	-1.82		.0	-.26		E
	D	0	1	1	-1.24		.1	-.22		D
	C	0	2	1	-.94	1.93	.3	-.28		C
	CD	0	2	1	-.37	1.54	.3	-.22		CD
	DE	0	11	6	.92	.57	1.1	-.19		DE
	AE	0	2	1	1.52	.26	.9	-.01		AE
	AB	0	1	1	1.55		.9	-.01		AB
	BD	0	31	18	1.72	.12	1.4	.03		BD
	BE	1	120	70	1.87	.06	1.0	.32		BE
28	Z	0	21	19	1.56	.22	1.3	-.17	26	Z
	A	1	91	81	2.00	.10	1.5	.17		A
	MISSING	***	611	85*	1.83	.04		-.03		
457	C	0	10	8	.45	.57	.5	-.38	369	C
	B	0	6	5	.90	.67	.7	-.19		B
	A	0	39	31	2.12	.08	1.5	.20		A
	D	1	70	56	1.90*	.11	2.8	.10		D
	MISSING	***	598	83*	1.85	.04		.02		

Figures 9 and 10 show the probability of receiving each score category as a function of ability (theta). The first graph shows a five-point item in which each score category is functioning as expected. As ability increases the probability of the next score point increases, and the probability of the previous score point decreases, and there is a well-defined threshold where adjacent score points are equally probable. The second figure shows an example where category probabilities are not well ordered – due to problems with one of the score categories. For this item, there is no point at which a four is the most likely score.

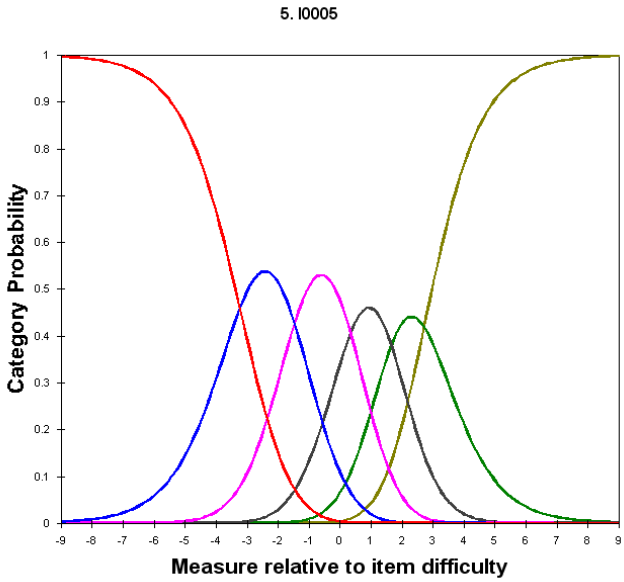

Figure 9. Polytomous Item With Ordered Score Categories

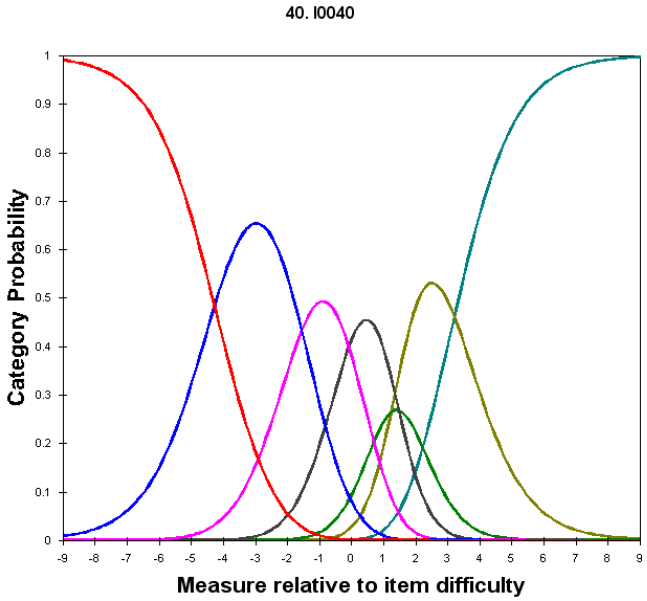

Figure 10. Polytomous Item With Problem Score Category

The decision to penalize incorrect options by not adding points vs. subtracting points for incorrect options will depend on the structure of an innovative item type. A multiple-answer multiple-choice item with unlimited selection (e.g., test takers can select all options) requires either dichotomous scoring or partial credit scoring with a mechanism for penalizing the selection of all options. Table 4 shows the descriptive statistics for dichotomous scoring, correct options, incorrect options, and scores with a penalty for incorrect options (partial credit). Dichotomous scoring clearly misses measurement opportunities and is highly vulnerable to any problems with the answer key. The subtraction of points based on incorrect choices both moderately improves the discrimination of the items and removes the risk of receiving high scores by selecting all options.

Table 4. Comparison point biserial correlations for different scoring options

ID	N	Dichotomous	# Correct	# Incorrect	Partial Credit
6	540	0.36	0.52	-0.41	0.59
7	540	0.41	0.53	-0.46	0.59
8	530	0.23	0.54	-0.37	0.56
12	354	-0.04	0.41	-0.2	0.45
13	359	0.32	0.46	-0.26	0.48
14	361	0.11	0.48	-0.37	0.54

One advantage of not restricting candidates to a limited number of options is the ability to rescore items. With an explicit instruction on the selection of options for hotspot or multiple-answer multiple-choice items (e.g,. “choose 2 options”) the ability to recover scores in the case of miskeyed items is limited. For example, item 12 in table X clearly has problems with at least one key (based on the dichotomous point biserial). Because test takers were not explicitly told to select 4 options, that item can be rescored without requiring the collection of new data.

The use of polytomously scored items with an IRT model requires careful evaluation of the scoring categories. Item analysis must include a review of response category frequency to check for unobserved or low-frequency options. Low-frequency score categories lead to inaccurate parameter estimates, while unobserved categories either indicate that the item is scored incorrectly or that it is too easy/difficult for the population. Items with unobserved or low frequency score categories may need to be discarded, or collateral information may be necessary to estimate IRT parameters.

While the analysis of responses and score categories on innovative items can indicate if an item is appropriately scored, it is important to ensure that scoring is done in a non- error-prone fashion. For an active testing program there should be a general scoring algorithm for each item type, not individually crafted rules for each item. While the "best" scoring option may vary slightly across items within an item type, the risk that complex scoring brings for correctly building, scoring, and analyzing tests in a large-scale operational setting by far outweigh the potential gains.

Conclusions

It is a clear measurement principle that examinee time should be spent on item content, not computer interface interpretation. Best practice for test and software development results in both innovative questions that can be easily interpreted and in fewer construct irrelevant user errors.

In this paper we have illustrated the benefit of several test development practices in the design of innovative item types. The first of these practices is the development of item writing materials such as item templates, item writing guidelines, and item writer training. The next process we addressed is usability testing. Conducting several rounds of usability tests with a representative sample of candidates early in the test development cycle can ensure that examinees will understand and be able to use the item type. Finally, we discussed item analysis methods that can be used to improve the quality of innovative item types.

References

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37(1),* 29-51.

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.

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.

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

Kayser, M., & Parshall, C.G. (2008, March). *Building a Global Innovative Test*. Presented at the annual meeting of ATP, Dallas, TX, March 2-5.

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

Linacre, J. M. (2006). WINSTEPS Rasch measurement computer program – version 3.64.2 [Computer software]. Chicago: Winsteps.com.

Masters, G. N. (1982) A Rasch model for partial credit scoring. *Psychometrika, 47(2),* 149-174.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16(2),* 159-176.

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

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

Parshall, C. G. & Harmes, J. C. (2008b). *Stages in designing innovative item types*. Presented at the annual meeting of ATP, Dallas, TX, March 2-5.

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.

Pommerich, M., & Burden, T. (2000, April). *From simulation to application: Examinees react to computerized testing.* Paper presented at the annual meeting of NCME, New Orleans.

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

Yen, W. M., & Fitzpatrick, A. R. (2008). Item response theory. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 111-153). Washington, DC: American Council on Education.

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