

Strategies and Processes for
**Developing Innovative Items
in Large-Scale Assessments**

Ellen Strain-Seymour, Ph.D.
Walter D. Way, Ph.D.
Robert P. Dolan, Ph.D.

PEARSON

The Pearson logo consists of the word "PEARSON" in a bold, serif font. Below the text is a horizontal line that is slightly curved, resembling a smile or a bridge.

June 2009

Introduction

Innovative items—computer-delivered items with interactivity that makes them unlike traditional, paper-and-pencil-delivered items—offer many benefits to students, teachers, and school systems (Parshall, Davey, and Pashley, 2000). These items have shown great promise for improving the quality, validity, and reliability of state large-scale assessments (Kane, 1992; Zenisky and Sireci, 2002). Innovative items are more engaging for students, can test students to a greater depth of knowledge, align more closely with curricular approaches, and permit more advanced applications of universal design and accessibility.

In addition, construct validity can be improved by decreasing reliance on construct-irrelevant knowledge and skills, such as using visuals to reduce reading load on non-reading tests. Furthermore, student interactions can be captured for rich, detailed analysis. Finally, innovative items can be practical from both state and local education system perspectives because they can be designed for automated scoring, avoiding the time and expense of hand-scoring. As such, they can represent the best of both worlds of selected response and constructed response formats. For all of these reasons, there is currently great enthusiasm for the potential use of innovative items in large-scale testing programs.

While the potential benefits of innovative items are great, the cost, effort, and time associated with their development is a significant challenge. To fully realize these benefits, more systematic approaches to developing and implementing innovative items need to emerge. Such plans can rely on efficient item production strategies to both overcome the obstacle of high item development costs and assure a high degree of usability.

In this paper we describe processes for developing high-quality innovative items in the context of large-scale assessments. We begin by defining what we mean by innovative items and further describing their benefits based on research and experiences from collaborations with our clients. Next, we discuss strategies and approaches for innovative item development, including the generation of templates for particular item types and processes by which multiple items might be generated from a single template. Finally, we describe the different roles of various experts that may contribute to developing innovative items and suggest a flow of steps through which innovative items are reviewed, refined, and prepared for administration. Through these steps we pay particular attention to the concept of usability and the goal of generating items that are accessible to all students.

Innovative Items: Definitions and Benefits

Defining Innovative Items

While working definitions of what constitutes an innovative item tend to vary, it is generally agreed that innovative items are computer-based test items that can not be easily translated to paper (Parshall, Davey, and Pashley, 2000). Some researchers classify innovative items as a subset of constructed response item types. Constructed response items, whether essay, short answer, or innovative, are distinguished from multiple-choice item types in that a student must construct a response by graphing, dragging and dropping, or typing, rather than by choosing one of a limited set of answer choices. A more liberal definition suggests that an

innovative item involves some kind of performance or interaction by way of responding or, in some cases, before responding, regardless of whether that response is “constructed” or multiple-choice. Through this logic, an innovative item could be as simple as a multiple-choice item that requires first viewing a set of slides through a microscope, or as elaborate as a virtual lab in which proper technique, calculation skills, lab step order, and results analysis are measured.

The Benefits of Innovative Items: A Research Review

The research around innovative items and the anecdotal evidence of enthusiastic responses to innovative items by students, teachers, and curriculum experts underscore these items’ many advantages:

- Measurement of a broader range of skills
- Increased authenticity
- Improved presentation of complex and dynamic information
- Reduced reading load
- Increased student engagement
- Reduced effect of successful guessing
- Reduced demands on working memory, allowing for more valid measurement
- Measurement of process skills and higher-order thinking

The most frequently cited justification for innovative items is their potential to measure skills that are not easily assessed through multiple-choice items. Such skills include higher-level cognitive skills, process skills, and complex problem-solving abilities. Huff and Sireci (2001) have suggested that higher-order skills such as reasoning, synthesis, and evaluation may be more effectively assessed by innovative items than by multiple-choice items. In addition, innovative item formats have the potential to provide “broader measurement of a construct domain as well as more efficient measurement of higher-level cognitive skills” (Huff & Sireci, 2001, p. 17). This improved measurement potential derives in part from increased authenticity.

Interactive environments with real-world resemblance can elicit specific behaviors that demonstrate proficiency within a certain domain (Bennett, 1999). For instance, items that involve students working with graphic organizers, line graphs, simulated software, primary documents, or simulated lab equipment can be expected to deliver greater authenticity when those are the very same activities that students perform in the classroom to demonstrate their knowledge and skills.

Another factor that researchers have isolated within studies of innovative items is their presentation of complex information. The dynamic representation of information within innovative items’ simulations and interactive environments can provide the necessary framework and support to allow for a meaningful assessment of higher-order skills (Harlen & Deakin Crick, 2003). Other studies support the idea that test-takers can interact realistically with data of considerable complexity within a multimedia environment, thereby facilitating performance measurements that would be difficult to measure through more traditional assessment strategies (Ridgway & McCusker 2004). Gorin describes such innovative items as “more cognitively rich contexts that mirror the complexity of the real world,” and as such, they provide opportunities to observe and assess student behavior and reasoning in ways that are not permitted by other more static item types (Gorin 2006, p. 31).

Content complexity can be more manageable in an innovative item for several reasons. With innovative items, the complexity is likely to be delivered in a fashion more similar to the hands-on learning context of the classroom (e.g., observing a chemical reaction in an animation that resembles a lab investigation rather than reading a description and seeing a static image). Students can better assimilate content complexity when they take an active role in processing the information (e.g., triggering the chemical reaction by adding the solution and then measuring the volume of the reactant). And lastly, the interactive components within an innovative item can provide students with a way to model their thoughts while working through a problem, much like manipulatives used within math instruction.

A number of studies investigate innovative items' greater construct validity, efficiency, and ability to assess a broad range of skills. In a study of item formats such as drag-and-drop that require test-takers to illustrate their understanding of hierarchies and content relationships, Jodoin (2001) found that innovative item formats provided more information about student skills across all ability levels. In addition, Maughn and Mackenzie (2004) confirmed the use of a simulated microscope to assess both process skills and the successful exercise of content knowledge within biology items. Lastly, Kumar, White and Helgeson (1993) demonstrated that a computer-based application providing basic tools for balancing chemical equations enabled students to achieve a higher level of performance than solving equations on paper.

The advantage of the computer-based environment was found to be particularly significant for novice students. Although the majority of such studies have focused on science content, one study of social studies items found that requiring students to represent history knowledge diagrammatically revealed deficiencies in their understanding of historical causality that were not shown in a conventional test (Masterman & Sharples, 2002).

A number of other benefits have been attributed to innovative items:

- **Reduced reading load through visual supports.** The graphic, animated, and interactive elements within an innovative item may provide information through multiple modalities to better support student understanding of a context, thereby reducing the reading load that may be required by an equivalent paper item. (A heavy reading load may introduce construct irrelevant variance in tests that are not intended to assess reading skills.) Additionally, research suggests that the visual representation afforded by the use of innovative items support how students construct meaning from the presented content (Kumar, White & Helgeson, 1993).
- **Decreased demand on working memory.** Studies demonstrate that innovative items can better support performance by reflecting back to students their selections within visual representations of knowledge (e.g., concept maps, graphic organizers, diagrams). The demand on working memory is decreased by providing an "external memory" for the student (Kumar, White & Helgeson, 1993, p. 7).
- **Richer diagnostic information.** With the ability to record, analyze, and assess processes and problem-solving strategies, additional diagnostic information can be obtained and used by students and teachers (Klieme,

2000; Birenbaum & Tatsuoka, 1987). Innovative items offer the potential to “bridge the gap between testing and instruction” by reporting not only the students’ final performance but also what processes need improvement (Schacter et al., 1997).

- **Effect on instruction.** By expanding the range of skills being assessed, instruction may be improved by encouraging the teaching of process skills that may be essential to the curriculum (Bennett, 1993).
- **Increased validity through the elimination of successful guessing.** Construct validity may be increased through the use of innovative items and constructed response items by reducing the impact of content-irrelevant test-taking skills and guessing associated with multiple-choice items (Huff & Sireci, 2001).
- **Better predictiveness.** Research suggests that expanded item types may be better predictors of educational performance (Frederiksen & Ward, 1978).
- **More student performance information.** Research from licensure has shown that innovative items take more time to complete than multiple-choice items, but they also provide more information (Jodoin, 2003).

Student and Educator Reactions to Innovative Items

As part of Pearson’s ongoing innovative item research, evidence of positive reactions from educators and test-takers has been collected in a variety of ways. For instance, one state testing program includes a student survey at the end of each online test, which includes an open-ended question inviting test-takers to provide more detail about aspects of the online test-taking experience that they may have disliked or enjoyed. Many of the student comments mirror the research, although often stated in less academic terms. For instance, students’ use of such descriptors as “fun,” “cool,” and “awesome” are suggestive of innovative items’ engaging delivery. Supporting the idea of authenticity, student feedback has included comments such as “I could see what I have learned happen right in front of me,” “reminds me of class,” and “it felt like you were really doing it [the lab].”

Students also comment on the role of innovative items’ visual and interactive support in promoting understanding, providing relief from heavy reading loads, and allowing engagement with complex content:

- *“I liked the video animation because usually I have to make a mental animation in my head.”*
- *“I really like being able to watch the videos and animations, and to work the interactive questions. They make it possible to test your knowledge of science in another way, and they are fun. Also they let you understand more than just illustrations.”*
- *“I enjoyed the one rock lab on one of the end questions. It really opened my eyes instead of reading a paragraph.”*
- *“I really liked the animated questions on this test because [...] I could visualize them and that helped me very much.”*

The standard item development process also involves evaluation of items by teachers, curriculum specialists, and content area experts. When innovative items are reviewed by these individuals, who are closest to the learning context, the following reactions have been noted as typical:

- Curriculum specialists are drawn to items that use innovative items to test core, multi-step processes. More granular assessment of the parts that comprise the whole is possible without removing context or complexity.
- Educators note, with enthusiasm, strong continuities between classroom activities and innovative items.
- Teachers comment on the transition to online testing in more overwhelmingly positive terms when innovative items are involved.

Innovative Items: Process and Development

Introducing Innovative Item Types

Introducing innovative items into an existing or new testing program begins with the identification of a student population or subject area that could benefit from innovative items' enhanced measurement capabilities. For tests that are already delivered exclusively or almost exclusively online, the logistics of expanding the test to include innovative item types are relatively easy to navigate. However, for testing programs with large proportions of paper test-takers, several options are possible.

In many cases, inadequate technology infrastructure in districts has slowed the transition to online testing. Acquiring sufficient numbers of computers to support all tests being administered to all students on a single day can be an expensive proposition. And interim solutions involving the administration of tests both on paper and online can be logistically complex (e.g., quality control of tests in different formats, accounting for mode effects, supporting districts for both types of test administration). Additionally, dual-mode testing often provides limited incentive for schools and districts to move to online testing.

One strategy for overcoming these barriers is to narrow the implementation strategy to a limited domain. For example, administering a single subject online and including innovative items in that test exposes all districts to the benefits of online testing and lessens the technology barrier by reducing the number of students who require access to a computer during the testing window. Equally important is the fact that this implementation strategy provides a compelling rationale for the transition.

The enthusiasm that students and teachers have for innovative items can be harnessed to achieve greater acceptance of online testing by districts, legislators, and funding bodies. Even with only a few innovative items per form, the potential of online testing to engage students and assess otherwise hard to measure skills can be demonstrated. For those states currently involved in dual-mode testing as part of a gradual transition to online, any hastening of the transition to online as the dominant test delivery mode, for one or more subjects, can reduce costs by decreasing the duration of the transition period.

Choosing a Venue for Innovative Items

Choosing a subject or program as the first to be expanded to include innovative items can be based on a number of criteria. Certainly logistical factors can play role: tests that are already online; tests with lower security requirements that can be delivered over a longer testing window to handle computer shortages; tests administered to a smaller proportion of students; or tests administered at the high school level (where student-to-computer ratios are closer to 1:1) can all be attractive options. Equally if not more important, however, is the opportunity to better serve a segment of the student population by improving the measurement potential of a given instrument.

Science is often the obvious first choice due to several factors. The emphasis on performance-based testing in science is a strong match to the potential within innovative items to create simulations and emulate lab environments. Science education's emphasis on context, observation, and visual information can be well served by the highly graphic nature of innovative items and their inclusion of time-based media such as sound, video, and animation. Process and problem-solving skills critical to science can also be assessed in a more robust fashion with innovative items.

Also, any tests with a college readiness component focused on process skills and higher-order thinking can be strong candidates for innovative items. For instance, more precise knowledge measurement and insight into process skills may be possible within an interactive math item that allows a student to demonstrate mastery over interim steps regardless of a correct final outcome.

Some subjects, however, are often over looked when it comes to interactive items. For instance, the social studies classroom can be rich with primary source usage, investigations of history through place and artifacts, the creation of presentation materials, analysis of economic and technological factors in the growth of nations, understandings of the effect of natural resources on civilizations, and visual analysis of posters and other media. Social studies assessments can involve greater authenticity and engagement through innovative items that mimic these activities in an assessment context. Similar arguments could also be made for English Language Arts (ELA) with the greater role of 21st century skills in ELA classrooms that engage in research, media literacy, and visual analysis.

For some, the discussion of innovative items conjures up thoughts of high-achieving students working through complex physics simulations, for instance, and producing responses that require sophisticated scoring algorithms. However, another way to think about innovative items is to consider the additional interactivity they offer as a support or accommodation provided to all or some students to help level the playing field. In such a view, innovative items could provide students with a non-traditional response mechanism (such as drag-and-drop) or give students an interactive stimulus within a multiple-choice item. Such a stimulus might provide supports for students to problem solve, model their thoughts, or visualize variations, without expanding the burden on scoring systems. For example, a math item might include pie slices to aid students in working through the addition of fractions. Or, instead of a revising-and-editing ELA item that requires students to (1) hold in working memory an image of paragraph 2 and 4 switched and (2) compare it to the effect of moving paragraph 5 to the end, an innovative item might allow students to move the

paragraphs around themselves to arrive at the best flow before making an answer selection.

Evidence-centered design provides a framework for constructing tests in terms of evidentiary arguments that is particularly agreeable to innovative items development (Mislevy and Haertel, 2006; Mislevy, Steinberg & Almond, 2003). This approach will further our ability to design innovative items with increased construct validity effectively and efficiently. Pearson is currently engaged in a research project funded by the National Science Foundation to apply evidence-centered design processes to the design, development, and implementation of technology-based science assessment tasks that will be included in a statewide science assessment (Fulkerson et al., 2009).

Templates and Item Types

Once a test, subject, or population has been identified as a potential opportunity for assessment enhancement through innovative items, a rationalized production method must be used to manage costs. The key to achieving sustainable innovative item development hinges on the ideas of template-based development, modularity, and code reuse. These ideas are important to reduce programming costs, save time, and allow content experts to manipulate innovative items without requiring technical assistance. The concept of item templates builds upon various frameworks and principles that have been proposed for understanding the different types of innovative items that can be developed (Dolan, Rose, Burling, Harms and Way, 2007; Scalise and Gifford, 2006; Zenisky and Sireci, 2002).

Templates are defined as reusable models or patterns used for creating individual instances of objects, such as test items. This approach better secures the affordability and reliability of the tasks and exercises developed for an online administration, making possible the goal of including innovative items in operational assessments in an efficient and sustainable manner.

Typically, an item template is associated with a single item type, although the template may contain a number of customizable parameters that allow the item type to take on any number of different "flavors" to match different assessment needs. In order to understand the potential role of templates in innovative item development, consider a continuum of innovative item types and templates, ranging from basic multiple-choice items, to items that can be delivered through content-neutral, reusable templates, to more customized, content-specific templates for which reuse may be difficult or impossible. This continuum is depicted in Figure 1.

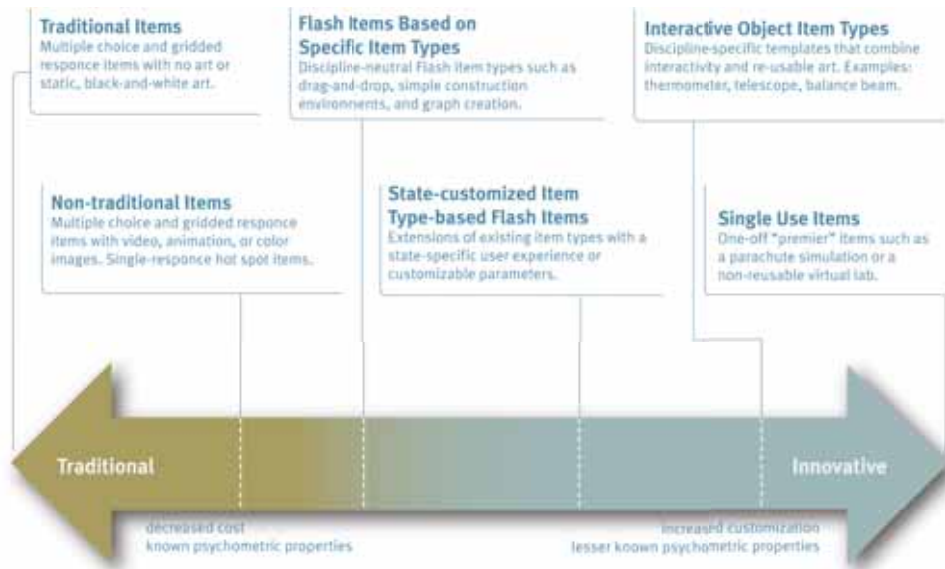


Figure 1. Continuum of Innovative Item Development.

As shown in Figure 1, templates can be used to support the development of a range of innovative item types. This flexibility makes this approach appropriate for creating items to support an assortment of test specifications across multiple states, as a function of factors such as task complexity, psychometric properties, cost, and ownership of the item delivery mechanism.

Furthering Template Development

Template development can be implemented conveniently using Adobe Flash®, a powerful software platform for creating rich multimedia, which is widely adopted in a variety of educational and non-educational settings. Template development within Flash® maximizes code reuse, minimizes maintenance efforts for items in an item bank, and reduces the required amount of effort for thorough quality control.

The flexibility and scalability afforded by templates is realized through the use of reusable interactive item elements that serve as components for building complete innovative assessment items. Item elements include the code, graphics, and content authoring tools associated with an element that might be used in multiple item templates. An item template, therefore, can be thought of as a collection of item elements that enable the creation of an entire item—including the embedding of sufficient data for scoring purposes—into which content is woven.

To build a robust pool of innovative items, it is necessary to have a full set of interactive item elements. For example, within a mathematics assessment project, the majority of item elements would be objects with math-specific behaviors. Thus, a rendering tool for coordinate grids might be an item element with considerable value across several math item types. This tool could be used by content developers to develop static graphics, or it could be made available to students to construct their own coordinate grid and other graphed elements for use within an item requiring a figural response. Alternatively, within a science assessment project, item type templates might be composed from a set of virtual lab equipment. A well-designed pool of item elements not only enables the efficient creation of a wide array of item

templates, it also provides students with a consistent user experience across assessments, despite the diversity of item types.

Usable, accessible design can—and should—begin at the item element level, per principles of universal design (Dolan and Hall, 2001; Ketterlin-Geller, 2005; Thompson, Johnstone, and Thurlow, 2002). For example, any text and functionality embedded within an item element must be able to operate within a larger strategy for accessibility by English language learners and by students both with and without disabilities, such as through a mouseless operation of buttons and controls or with a Mathematical Markup Language (MathML)-aware text-to-speech engine.

Each item element should also conform to standard rules that together define a uniform and intuitive user experience. In this regard, template-based item production contributes to construct validity. When students are exposed to a range of typical interaction types via tutorials and when those interaction types are submitted to rigorous accessibility and usability testing, the potential for construct irrelevant variance, that can be introduced through varying computer expertise, can be minimized.

Templates and Content Authoring

Templates are based on a set of rules that define the item type and a set of parameters that allow item authors to customize the items generated from the templates. Templates should also have embedded options for specific styles that might be preferred by particular states for their assessment programs. Template-based test development begins with content-based concepts that are usually articulated and sketched out using various prototyping strategies, such as storyboarding. Once a concept is originated, three tasks are necessary for template-based test development:

1. The programming of the template and the content authoring tool that accompanies the template, handled by a Flash® programmer.
2. The creation of art according to template specifications, handled by a graphic designer or animator.
3. The authoring of content, the selection of parameter values to customize the template, and the previewing of the outcome, handled by item writers and content specialists.

The differentiation between the first and last of these tasks is significant, since the creation of the template requires high-level programming skills whereas non-technical content experts are able to do the Flash® tasks associated with content coordination.

A number of mechanisms can be explored to streamline item creation and make content creation tools accessible to non-technical item authors. Prior experience with innovative item development has shown that high-quality item development is best facilitated by putting item creation tools in the hands of the users who best understand the content. Similarly, item review by experts and educators, and bias review boards is most effective when reviewers can interact with the item just as the student would, and make and evaluate edits immediately, within the context of the item. Therefore, tools for creating, editing, and reviewing items should not require extensive technical knowledge and should include robust editing and previewing capabilities.

Item Type Development

Arriving at a proposed set of item types involves a compromise between supporting item variability and adhering to a pattern. This process begins with the identification and analysis of the following:

- Knowledge and skill areas that are difficult to measure with traditional item types
- Software and online tools that are commonly used as part of classroom instruction
- Physical tools (e.g., protractor, compass, ruler, lab equipment) that are used in the classroom, which may or may not be used during paper-based assessments
- Specific areas where higher-order thinking skills are required within a subject area and the classroom exercises that tend to engage and develop these skills
- Concepts and problem-solving skills that are best assessed within the context of real-world situations

This process should produce a series of item type ideas that effectively match the benefits of interactive items (simulation of context, virtual equivalents of physical tools, real-world scenarios without increased reading load, robust response mechanisms) with the following goals:

- Measurement of higher order thinking skills
- Authenticity through continuity with instructional tools
- Increased student engagement
- Reduction of construct-irrelevant variance caused by reading load and language barriers
- Greater validity

Further exploration of item type ideas should lead to the general contours of an item type template, including: a set of rules for student interaction, parameters that determine content variability, and an initial visualization, which may include the alignment of certain screen zones to different types of content. For instance, in a biology test, a microscope might be deemed a tool of central importance. It can provide context to aid identification as a student concretely engages with the issue of image scale by selecting an ideal magnification level. It can also provide a mechanism for comparing different types of specimens using an authentic context. Lastly, it can also be used to provide information on whether a student is employing proper technique.

The assessment goals for equipment inclusion should guide the formulation of a set of rules for student interaction. For example, with the microscope, the following interactions could be established:

- The student should be able to view a single slide or view multiple slides.
- Clicking a slide should make that slide viewable through the microscope. (In the case of only one available slide, the student will not need to click the slide; it will automatically be in view.)
- The student should be able to choose a microscope objective in order to view the slide at a particular magnification level.
- Focusing using coarse focus at the lowest magnification level should be necessary to viewing the slide, with the aid of fine focus, at the higher magnification levels, in the same way as with an actual microscope.

- Certain functionality pertaining to a microscope will not be necessary, for example: plugging the microscope in, turning it on, turning on the light, fixing the slide in place using the slide clips, cleaning the lens, and adjusting the eye piece.

The parameters for content variability also need to be determined and built into the template. For instance, a set of content authoring rules might be applied to the microscope:

- The use of the microscope can be combined with multiple response mechanisms such as multiple-choice and numeric constructed response.
- Between one and four slides may be offered within any item.
- Each slide can have its own label (e.g. Epithelial Cells) or can default to a generic labeling scheme (e.g. Slide 1).
- The image shown in each slide can be either a static image or a video.
- The item can have between one and three magnification levels available with all of the slides. Those magnification levels can be specified at the item level.
- Coarse focus can be disabled above a specified magnification level.
- A permission line or credit can be displayed for each slide.


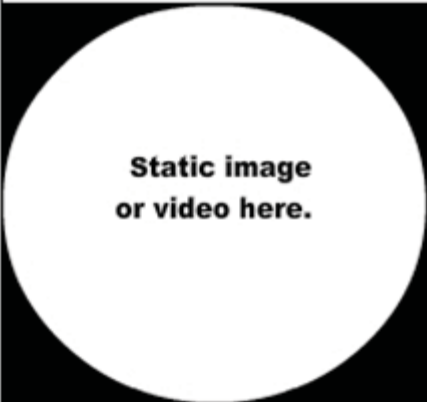
Introductory text determined by item author. Auto-centered vertically		
Click to choose a slide		Lab notebook, Multiple Choice, Or Grid-in
SLIDE A		
SLIDE B		
SLIDE C		
		

Figure 2: Example of an early storyboard of the microscope item template

Observe each of the four prepared slides by clicking on the slide label.
Use the appropriate focus controls to bring the cells into focus.


Click to choose a slide


Slide 1

Slide 2

Slide 3

Slide 4





Which two slides provide the best illustration of the differences between plant and animal cells?

A Slides 1 and 2

B Slides 2 and 3

C Slides 1 and 3

D Slides 2 and 4

Figure 3: An item created from the microscope template

A second example involves innovative geometry items. The item type idea generation phase might have led to the identification of a series of student expectations focused on figure transformation, and the desire to provide functionality that is similar to a commonly used instructional software package for the manipulation of geometric figures. In this case, the content variability may be guided by the following rules:

- The response mechanism may be either numeric constructed response or multiple-choice.
- A transformation palette appearing in the stem will be one of four types: rotation, dilation, translation, or reflection.
- The initial shape or line can be determined by the content creator.
- Any limitations on how the student can rotate, dilate, translate, or reflect the initial shape can be specified by the content creator.

Following the creation of the template and the content authoring widget, any number of items can be created for very little cost. Figures 4 and 5 show examples of template-based items.

Line AV is the median of triangle ABC shown below. Use the reflection tool below to reflect the triangle in different ways.

Reflection Tool
Line of Reflection

x-axis $y = x$
 y-axis $y = -x$

Reflect Shape

Which of the following is true?

A The reflected triangle is always in a different quadrant than the original triangle.

B Line $A'V'$ is always either perpendicular or parallel to the x -axis.

C The measure of angle B changes depending on the line of reflection.

D Triangle ABC is a right triangle, but triangle $A'B'C'$ is not always a right triangle.

Figure 4: Example of an innovative geometry item using the transformation palette.

The three medians of a triangle are shown on the coordinate grid below. Clicking on and dragging any vertex of the original triangle will transform the triangle.

Which of the following statements is true in all transformations performed?

A No two medians are ever perpendicular.

B One of the medians of an isosceles triangle is an altitude of the triangle.

C The medians of a triangle always form six right triangles.

D The medians of a scalene triangle do not always intersect.

Figure 5: An innovative geometry item using the dynamic polygon template.

Innovative Item Development Process Flow

Once item type ideas have been generated and the template rules and parameters have been documented, development should proceed in such a way as to validate the templates against the items and vice versa. Figure 6 presents sample development and review milestones that demonstrate this interplay of template and item development. As shown, the creation of item type templates and the items themselves are parallel and linked activities, requiring a balance between rules and flexibility. This provides a mechanism for creating multiple items from each new item type to help distinguish challenges with the item type from any item-specific issues that may arise.

Throughout these parallel development streams, review steps are built in to provide adequate guidance, oversight, and collaborative input for the development of item type templates and items. These review and evaluation activities, related to the items, the templates, and the processes, typically include storyboard reviews by content experts, usability testing, and accessibility, universal design, and bias reviews. A snapshot view of the types of innovative item review processes that would typically occur between Pearson and our clients is also provided.

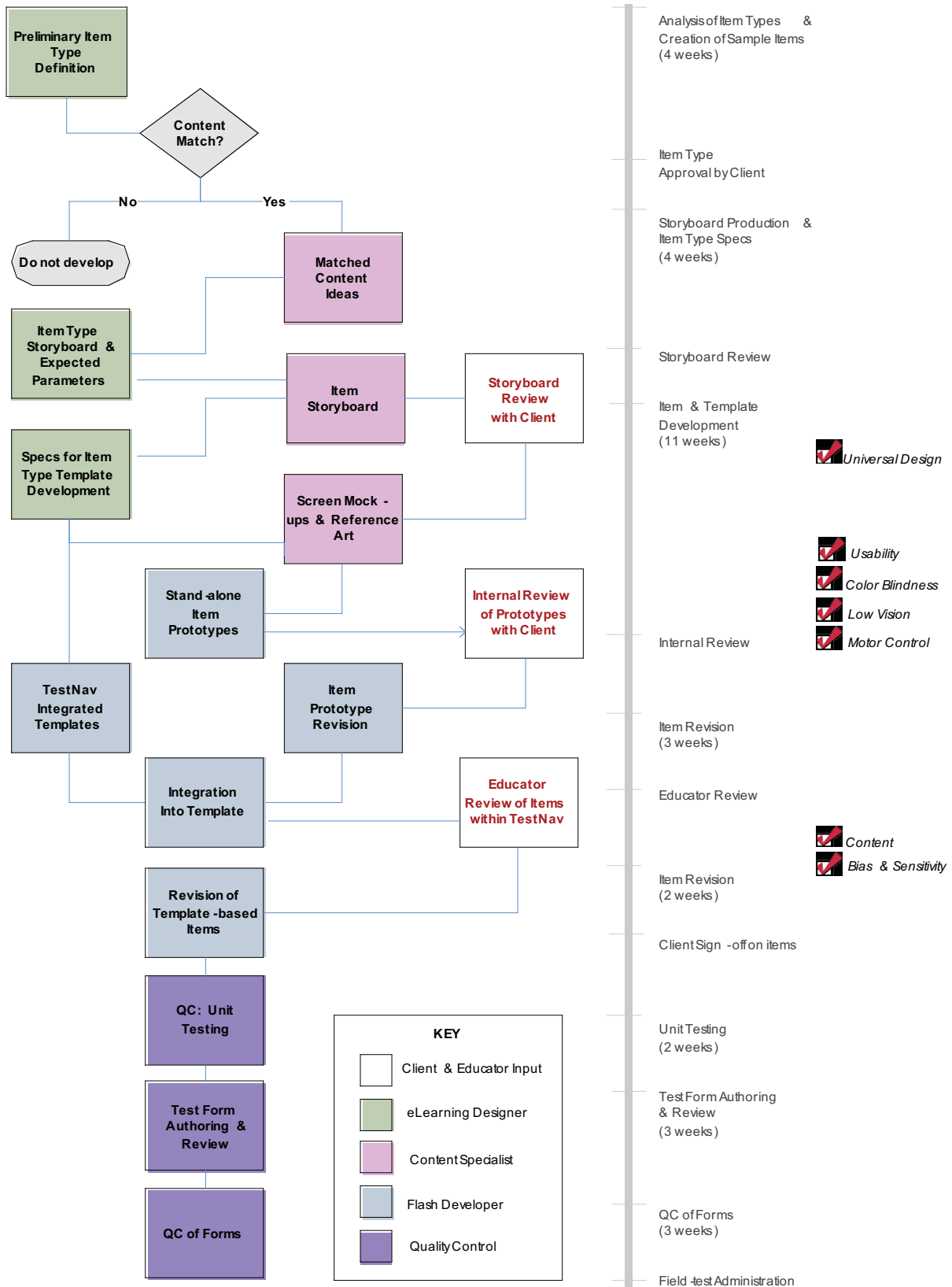


Figure 6. Sample Development and Review Milestones for Innovative Items.

Usability and Accessibility

In contrast to traditional items, the design and development of innovative items must consider that user interfaces are more complex and might be unfamiliar to students. If these factors are not adequately controlled for, increased construct-irrelevant variance may be introduced into student results. This situation can be addressed by following a user-centered design approach (Norman, 1988), in which the development process draws on user input at multiple instances during development to validate design ideas and correct any usability problems. Adherence to user-centered design throughout the development process will help assure that the interactivity within any item type is easily and immediately understood by test-takers, without requiring time or energy to learn how to operate the interface.

While user input may be sought in informal ways at design crossroads, more formal usability studies may be appropriate at different points within the development cycle. For example, a usability study may be undertaken using rough or “low-fidelity” prototypes, such as designs sketched out on paper, prior to a significant development investment, and a late-stage study using higher fidelity prototypes, such as partially interactive designs, might occur once templates and items are close to completion.

Whenever possible, usability studies should involve subjects that represent the target population of the assessment. One model of a usability study consists of one-on-one sessions with students using a think-aloud/cognitive lab protocol to foster an understanding of subjects’ cognitive processes as they step through the prototypes. Screen captures and audio transcriptions can be analyzed with findings documented in a formal report. When done at an early development phase, any design missteps can be quickly identified and corrected before a significant investment is made in software development efforts.

Additional usability studies at subsequent stages within development can be conducted remotely with the primary data consisting of screen captures of test-takers interacting with sample items that illustrate the particular item types. Additional survey data may be collected as well. Any usability glitches discovered at this stage can be corrected as the items are finalized for field testing. Sufficient time should be provided in this phase to do limited follow-up, or iterative usability studies, if warranted.

In order for innovative items to be fair and accessible for a wide range of students, including those with disabilities and who are English language learners, the software industry’s best practices in the area of usability engineering should be combined with universal design standards established for assessment (Dolan and Hall, 2001; Ketterlin-Geller, 2005; Thompson, Johnstone, and Thurlow, 2002). To facilitate this process, Pearson has developed the *Universal Design for Computer-based Test Guidelines* (Dolan, Burling, Harms, Beck, Hanna and Jude, 2006; Dolan, Rose, Burling, Harms, and Way, 2007), which include special considerations for the rich media and interactivity environment of innovative items. These considerations are made not only to incorporate testing accommodations for students with disabilities, but to help decrease construct-irrelevance for all students.

In addition to the usability studies and checks against universal design principles, Pearson includes several additional reviews as part of the innovative item development process:

- Use of tools such as IBM's aDesigner™ tool to simulate three types of color blindness to evaluate comprehensibility of items for color blind students.
- An analysis of possible mechanisms within each item type for enlarging, increasing contrast, and providing alternate text for low-vision students.
- A "mouseless" test to assure that the items can be used by students with physical disabilities that prevent their use of a mouse, such as adequate keyboard shortcuts or alternative ways to move objects within drag-and-drop environments.

Although it may not be possible to make all innovative items completely accessible for all students, Pearson is committed to making all reasonable efforts in this area and to better understand the challenges and costs of supporting various types of accommodations within innovative items. In cases where items fail to pass all of the accessibility reviews, it is useful to document the failings and research ways to address these issues in future iterations of item type designs.

Summary and Conclusions

This paper addressed strategies and processes for developing innovative items for large-scale assessments. Clearly, innovative items offer multiple benefits to students, teachers, and schools. In particular, innovative items have the promise of better evaluating the cognitive, process, and problem-solving skills that will be critical to students' success in the 21st century workforce.

A primary approach described in this paper is template-based item development, which can provide an affordable and reliable way for states to create and sustain pools of innovative items. By designing templates for innovative item types, developers can reduce programming costs and save development time. In addition, by providing content experts with the tools they need to customize and edit templates, and preview and interact with content in the same format students will use, developers can reduce the time and effort associated with programming and content creation.

A user-centered design approach can enhance the development of innovative items and help focus efforts on producing items that are accessible to all test-takers. Furthermore, a development process that includes user input at multiple points serves to validate design ideas and correct any usability problems before testing begins. Such an approach is in accordance with Universal Design and helps to minimize the influence of construct-irrelevant factors on test performance.

It seems clear that the use of innovative items is merely a first step in the next generation of the assessment industry. This paper provides guidance and suggestions for taking this step. As experience is gained and school technology infrastructures advance, the full benefits of technology-based assessment will be realized, and innovative items will become so widespread that they will no longer be considered "innovative".

References

- Bennett, R. E. (1993). On the meanings of constructed response. In R. E. Bennett & W. C. Ward (Eds.), *Construction vs. choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 1-27). Hillsdale, NJ: Erlbaum.
- Bennett, R. E., Steffen, M., Singley, M. K., Morley, M., Jacquemin, D. (1997). Evaluating an automatically scorable, open-ended response time for measuring mathematics reasoning in computer-adaptive tests. *Journal of Educational Measurement, 34*(2), 162-176.
- Bennett, R. E. (1999). Using new technology to improve assessment. *Educational Measurement: Issues and Practice, 18*, 5-12.
- Birenbaum, M., & Tatsuoka, K. K. (1987). Open-ended versus multiple-choice response formats – It does make a difference for diagnostic purposes. *Applied Psychological Measurement, 11*, 385-395.
- Dolan, R. P., & Hall, T. E. (2001). Universal Design for Learning: Implications for large-scale assessment. *IDA Perspectives, 27*(4), 22-25.
- Dolan, R. P., Burling, K. S., Harms, M., Beck, R., Hanna, E., Jude, J., et al. (2006). *Universal Design for Computer-Based Testing Guidelines*. Retrieved May 4, 2009, from <http://www.pearsonedmeasurement.com/cast/index.html>.
- Dolan, R. P., Rose, D. H., Burling, K. S., Harms, M., & Way, W. (2007). *The Universal Design for Computer-Based Testing Framework: A structure for developing guidelines for constructing innovative computer-administered tests*. Paper presented at the National Council on Measurement in Education Annual Meeting, Chicago, IL.
- Frederiksen, N., & Ward, W. C. (1978). Measures for the study of creativity in scientific problem solving. *Applied Psychological Measurement, 2*, 1-24.
- Fulkerson, D., Nichols, P. D., Mislavy, R. J., Liu, M., Zalles, D. R., Fried, R. C., Villalba, S. E., Debarger, A. H., Cheng, B., Mitman, A. L., Haertel, G. D. & Cho, Y. (2009). *Leveraging Evidence-Centered Design (ECD) Within Scenario-Based Statewide Science Assessment: Research Findings: Leveraging ECD in Scenario-Based Science Assessments*. Paper presented at the Annual Meeting of the American Educational Research Association. San Diego, CA, USA: April 13-17.
- Gorin, J. S. (2006). Test Design with Cognition in Mind. *Educational Measurement: Issues and Practice, 25*, 21-35.
- Harlen, W. & Deakin Crick, R., (2003). A systematic review of the impact on students and teachers of the use of ICT for assessment of creative and critical thinking skills. Evidence for Policy and Practice Co-ordinating Centre Department for Education and Skills, London.
- Huff, K. L, & Sireci, S. G. (2001). Validity Issues in Computer-Based Testing. *Educational Measurement: Issues and Practice, 20*, 16-25.

- Jodoin, M. G. (2001). *An empirical examination of IRT information for innovative item formats in a computer-based certification testing program* (Laboratory of Psychometric and Evaluative Methods Research Report No. 417). Amherst, MA: School of Education, University of Massachusetts.
- Jodoin, M. G. (2003). Measurement efficiency of innovative item formats in computer-based testing. *Journal of Educational Measurement, 40*, 1-15.
- Kane, M.T. (1992). The assessment of professional competence. *Evaluation and the Health Professions, 15*, 163-182.
- Ketterlin-Geller, L. R. (2005). Knowing What All Students Know: Procedures for Developing Universal Design for Assessment. *Journal of Technology, Learning, and Assessment, 4*(2), 1-23.
- Klieme E. (2000). *Assessment of Cross-Curricular Problem-Solving Competencies*. Berlin, Germany: Max-Planck Institute for Human Development, Centre for Educational Research.
- Kumar D. D., White A.L., Helgeson S. L. (1993). Effect of HyperCard and traditional performance assessment methods on expert-novice chemistry problem-solving. Paper presented at the Annual Meeting of the National Association for Research in Science Teaching. Atlanta, GA, USA: April 15-19.
- Lewis, D. M., Mitzel, H. C., & Green, D. R. (1996). *Standard setting: A bookmark approach*. Paper presented at the CCSSO National Conference on Large Scale Assessment.
- Masterman L, Sharples M. (2002) A theory-informed framework for designing software to support reasoning about causation in history. *Computers and Education 38*,165-185.
- Maughan, S. & Mackenzie, D. (2004). *Bioscope: The Assessment of Process and Outcomes using the TRIAD System*. Proceedings of the 8th CAA Conference, Loughborough: Loughborough University.
- Mislevy, R. J. & Haertel, G. (2006). *Implications of Evidence-Centered Design for Educational Testing (PADI Technical Report 17)*. Menlo Park, CA: SRI International.
- Mislevy, R. J., Steinberg, L. S. & Almond, R. G. (2003). On the structure of educational assessment. *Measurement: Interdisciplinary Research and Perspective, 1*(1): 3-62.
- Norman, D. (1988). *The Design of Everyday Things*. New York, NY: Doubleday
- Parshall, C. G., Davey, T. & Pashley, P. (2000). Innovative item types for computerized testing. In W.J. van der Linden and C.A.W. Glas (eds.), *Computerized Adaptive Testing: Theory and Practice*. Netherlands: Kluwer Academic Publishers.
- Ridgway, J., McCusker, S., Pead, D. (2004). *Literature Review of E-assessment*. (Futurelab). Bristol, UK.

- Scalise, K., & Gifford, B. (2006). Computer-Based Assessment in E-Learning: A Framework for Constructing "Intermediate Constraint" Questions and Tasks for Technology Platforms. *Journal of Technology, Learning, and Assessment, 4*(6).
- Schacter J., Herl H.E., Chung G.K.W.K., O'Neil Jr. H.F., Dennis R.A., Lee J.J. (1997). *Feasibility of a web-based assessment of problem-solving*. Paper presented at the Annual Meeting of the American Educational Research Association. Chicago, IL, USA: March 24-28.
- Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large-scale assessments* (NCEO Synthesis Report 44). Minneapolis, MN: University of Minnesota, National Center on Education Outcomes.
- Zenisky, A. L., & Sireci, S. G. (2002). Technological innovations in large-scale assessment. *Applied Measurement in Education, 15*(4), 337-362.