J·T·L·A

The Journal of Technology, Learning, and Assessment

Volume 4, Number 6 · June 2006

Computer-Based Assessment in E-Learning: A Framework for Constructing "Intermediate Constraint" Questions and Tasks for Technology Platforms

Kathleen Scalise & Bernard Gifford

www.jtla.org

A publication of the Technology and Assessment Study Collaborative
Caroline A. & Peter S. Lynch School of Education, Boston College

J·T·L·A

Volume 4, Number 6

Computer-Based Assessment in E-Learning: A Framework for Constructing "Intermediate Constraint" Questions and Tasks for Technology Platforms

Kathleen Scalise & Bernard Gifford

Editor: Michael Russell
russelmh@bc.edu
Technology and Assessment Study Collaborative
Lynch School of Education, Boston College
Chestnut Hill, MA 02467

Copy Editor: Kevon R. Tucker-Seeley
Design: Thomas Hoffmann
Layout: Aimee Levy

JTLA is a free on-line journal, published by the Technology and Assessment Study Collaborative, Caroline A. & Peter S. Lynch School of Education, Boston College.

Preferred citation:

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). Retrieved [date] from http://www.jtla.org

J·T·L·A

Abstract:

Technology today offers many new opportunities for innovation in educational assessment through rich new assessment tasks and potentially powerful scoring, reporting and real-time feedback mechanisms. One potential limitation for realizing the benefits of computer-based assessment in both instructional assessment and large scale testing comes in designing questions and tasks with which computers can effectively interface (i.e., for scoring and score reporting purposes) while still gathering meaningful measurement evidence. This paper introduces a taxonomy or categorization of 28 innovative item types that may be useful in computer-based assessment. Organized along the degree of constraint on the respondent's options for answering or interacting with the assessment item or task, the proposed taxonomy describes a set of iconic item types termed "intermediate constraint" items. These item types have responses that fall somewhere between fully *constrained* responses (i.e., the conventional multiple-choice question), which can be far too limiting to tap much of the potential of new information technologies, and fully *constructed* responses (i.e. the traditional essay), which can be a challenge for computers to meaningfully analyze even with today's sophisticated tools. The 28 example types discussed in this paper are based on 7 categories of ordering involving successively decreasing response constraints from fully selected to fully constructed. Each category of constraint includes four iconic examples. The intended purpose of the proposed taxonomy is to provide a practical resource for assessment developers as well as a useful framework for the discussion of innovative assessment formats and uses in computer-based settings.

J·T·L·A

Computer-Based Assessment in E-Learning: A Framework for Constructing "Intermediate Constraint" Questions and Tasks for Technology Platforms

Kathleen Scalise[1]
University of Oregon
Bernard Gifford[2]
University of California, Berkeley

Introduction

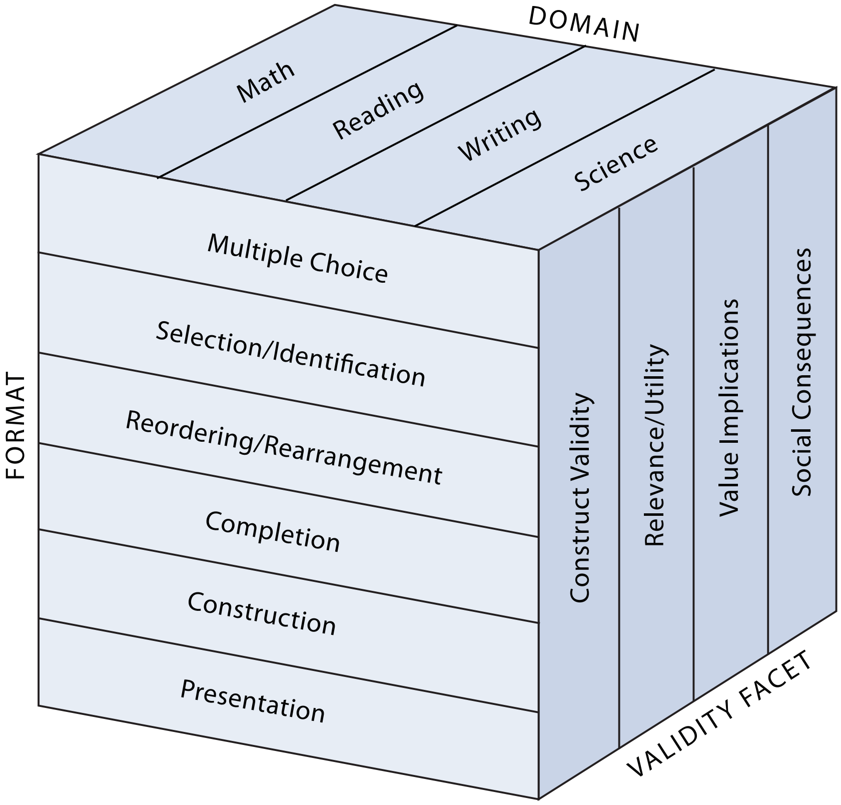
Computers and electronic technology today offer myriad ways to enrich educational assessment both in the classroom and in large-scale testing situations. With dynamic visuals, sound and user interactivity as well as adaptivity to individual test-takers and near real-time score reporting, computer-based assessment vastly expands testing possibilities beyond the limitations of traditional paper-and-pencil tests. Through these and other technological innovations, the computer-based platform offers the potential for high quality *formative* assessment that can closely match instructional activities and goals, make meaningful contributions to the classroom, and perhaps offer instructive comparisons with large-scale or summative tests. As the digital divide lessens, it would seem that technology should be poised to take advantage of these new frontiers for innovation in assessment, bringing forward rich new assessment tasks and potentially powerful scoring, reporting and real-time feedback mechanisms for use by teacher and students.

One potential limitation for realizing the benefits of computer-based assessment comes in designing questions and tasks with which computers can effectively interact, including scoring and score reporting, while still gathering meaningful measurement evidence. The question type currently dominating large-scale computer-based testing and many e-learning assessments is the standard multiple-choice question, which generally includes a prompt followed by a small set of responses from which students are expected to select the best choice. This kind of task is readily scorable by a variety of electronic means and offers some attractive features as an assessment format (see discussion below on multiple choice formats). However, if e-learning developers adopt this format alone as the focus of assessment formats in this emerging field, much of the computer platform's potential for rich and embedded assessment could be sacrificed.

According to some researchers, ubiquitous multiple-choice testing sometimes encourages "poor attitudes toward learning and incorrect inferences about its purposes…for example that there is only one right answer, that the right answer resides in the head of the teacher or test maker, and that the job of the student is to get the answer by guessing" (Bennett, 1993, p. 24). Movements toward authentic assessment, alternative assessment, performance assessment, dynamic assessment, portfolio systems, constructed response, higher-order assessment and other approaches favoring richer assessment tasks are often based on consequential validity arguments about deleterious effects on teaching and learning of narrow assessments in the classroom (Osterlind, 1998). Some cognitive theorists argue that the multiple-choice format presumes, often without sufficient basis, that complex skills can be decomposed and decontextualized. Moreover, some critics maintain that in practice, this format over-relies on well-structured problems with algorithmic solutions and that in theory, it builds on a view of learning that knowledge is additive rather than integrative of developing knowledge structures (Glaser, 1988, 1991; Resnick & Resnick, 1992; Shepard, 1991a, 1991b).

In this paper, we introduce a taxonomy or categorization of 28 innovative item types that may be useful in computer-based assessment. Organized along the degree of constraint on the respondent's options for answering or interacting with the assessment item or task, the proposed taxonomy (shown in Table 1, on page 9) describes a set of iconic item types termed "intermediate constraint" items. These item types have responses that fall somewhere between fully *constrained* responses (i.e., the conventional multiple-choice question), which can be far too limiting to tap much of the potential of new information technologies, and fully *constructed* responses (i.e. the traditional essay), which can be a challenge for computers to meaningfully analyze even with today's sophisticated tools. The 28 example types discussed in this paper are based on 7 categories of ordering involving successively decreasing response constraints from fully selected to fully constructed. Each category of constraint includes four iconic examples. References for the Taxonomy were drawn from a review of 44 papers and book chapters on item types and item designs – many of them classic references regarding particular item types – with the intent of consolidating considerations of item constraint for use in e-learning assessment designs.

Figure 1: Bennett's "Multi-faceted Organization Scheme" (Bennett, 1993, p. 47)[3]



Intermediate Constraint Taxonomy for E-Learning Assessment Questions & Tasks

Questions, tasks, activities and other methods of eliciting student responses are often called items in the assessment process[4]. In the computer-based platform, we argue that almost any type of interaction with a user can be considered an assessment item. Note that a working definition we propose for an assessment item (when speaking in reference to technology) is any interaction with a respondent from which data is collected with the intent of making an inference about the respondent.

Given this definition, there are many ways in which assessment items can be innovative when delivered by computer. One organizational scheme describes innovative features for computer-administered items, such as the technological enhancements of sound, graphics, animation, video or other new media incorporated into the item stem, response options or both (Parshall, Davey, & Pashley, 2000). But other classification possibilities are myriad, including how items function. For some innovative formats, students can, for instance, click on graphics, drag or move objects, re-order a series of statements or pictures, or construct a graph or other representation. Or the innovation may not be in any single item, but in how the items flow, as in branching through a changing series of items contingent on an examinee's responses.

Much of the literature on item types is concerned with innovations of the observation – the stimulus and response – that focus on the degree of construction versus selection, or constraint versus openness, in the response format. A number of characteristics are common to most constructed-response and performance formats:

> First and perhaps most obvious, these alternative formats require an examinee to supply, develop, perform, or create something. And, typically, these tasks attempt to be more engaging to the examinee than conventional multiple-choice items. Often, they employ real-world problems that people of a comparable age and peer status may encounter in daily life, such as asking school-age children to calculate from a grocery store purchase, or for high schoolers, to complete a driver's license application or examine an insurance policy. [They] are generally scored by comparing and contrasting an examinee's responses to some developed criteria, sometimes elucidated in lengthy descriptions called "rubrics" (Osterlind, 1998, p. 205).

So-called open-ended items cover a lot of territory, however, and organizing schemes for the degree of constraint and other measurement aspects regarding items can be helpful (Bennett, 1993). An organization by Bennett is shown in Figure 1 on the previous page.

Drawing on the concept of what might be called a *constraint dimension*, in this paper we develop and present the *Intermediate Constraint Taxonomy for E-Learning Assessment Questions and Tasks*, shown in Table 1 on the following page. The Taxonomy describes and gives examples of 28 iconic intermediate constraint (IC) item types that feature a variety of innovations in the stimulus and/or response of the observation. IC types may be useful, for instance, with automated scoring in computer-based testing (CBT). IC items and task designs are beginning to be used in CBT, with response outcomes that are promising for computers to readily and reliably score, while at the same time offering more freedom for the improvement of assessment design and the utilization of computer-mediated functionality. The taxonomy of constraint types described here includes some characteristics, previous uses, strengths and weaknesses of each type, and we present examples of each type in figures in this paper. Intermediate constraint tasks can be used alone for complex assessments or readily composited together, bundled and treated with bundle (testlet) measurement models (Scalise, 2004).

At one end of the spectrum, the most constrained selected response items require an examinee to select one choice from among a few alternatives, represented by the conventional multiple-choice item. At the other end of the spectrum, examinees are required to generate and present a physical performance under real or simulated conditions. Five intermediary classes fall between these two extremes in the Taxonomy and are classified as selection/identification, reordering/rearrangement, substitution/correction, completion, and construction types.

Note that all item types in the item taxonomy can include new response actions and media inclusion. Thus, by combining intermediate constraint types and varying the response and media inclusion, e-learning instructional designers can create a vast array of innovation assessment approaches and could arguably match assessment needs and evidence for many instructional design objectives.

Media inclusion, simulations, within-item interactivity and data-rich problem-solving in which access to rich resources such as books, resources and references are made available online, are all innovations that can be incorporated in many of the item types discussed below. Examples of media inclusion are numerous and include the multimedia rich National Center for Research on Evaluation, Standards, and Student Testing (CRESST) examples (Chung & Baker, 1997), simulations (Parshall, Spray, Kalohn, & Davey, 2002), and data-rich assessment and tracking of problem-solving paths such as those exemplified in Interactive Multimedia Exercises (IMMEX) (Pellegrino, Chudowsky, & Glaser, 2001). A discussion of the use of multimedia in large-scale computer-based testing programs by Bennett and co-authors considers the incorporation of dynamic stimuli such as audio, video and animation in history, physical education and science assessment (Bennett et al., 1999).

Table 1: Intermediate Constraint Taxonomy for E-Learning Assessment Questions and Tasks

This table shows twenty-eight item examples organized into a taxonomy based on the level of constraint in the item/task response format. The most constrained item types, at left in Column 1, use fully selected response formats. The least constrained item types, at right in Column 7, use fully constructed response formats. In between are "intermediate constraint items," which are organized with decreasing degrees of constraint from left to right. There is additional ordering that can be seen "within-type," where innovations tend to become increasingly complex from top to bottom when progressing down each column.

Most **Constrained** → ***Least*** **Constrained**

Fully Selected — *Intermediate Constraint Item Types* — *Fully Constructed*

Less **Complex** ↓ ***More*** **Complex**

1. Multiple Choice	**2. Selection/ Identification**	**3. Reordering/ Rearrangement**	**4. Substitution/ Correction**	**5. Completion**	**6. Construction**	**7. Presentation/ Portfolio**
1A. *True/False* (Haladyna, 1994c, p.54)	2A. *Multiple True/False* (Haladyna, 1994c, p.58)	3A. *Matching* (Osterlind, 1998, p.234; Haladyna, 1994c, p.50)	4A. *Interlinear* (Haladyna, 1994c, p.65)	5A. *Single Numerical Constructed* (Parshall et al, 2002, p. 87)	6A. *Open-Ended Multiple Choice* (Haladyna, 1994c, p.49)	7A. *Project* (Bennett, 1993, p.4)
1B. *Alternate Choice* (Haladyna, 1994c, p.53)	2B. *Yes/No with Explanation* (McDonald, 2002, p.110)	3B. *Categorizing* (Bennett, 1993, p.44)	4B. *Sore-Finger* (Haladyna, 1994c, p.67)	5B. *Short-Answer & Sentence Completion* (Osterlind, 1998, p.237)	6B. *Figural Constructed Response* (Parshall et al, 2002, p.87)	7B. *Demonstration, Experiment, Performance* (Bennett, 1993, p.45)
1C. *Conventional* or *Standard Multiple Choice* (Haladyna, 1994c, p.47)	2C. *Multiple Answer* (Parshall et al, 2002, p.2; Haladyna, 1994c, p.60)	3C. *Ranking & Sequencing* (Parshall et al, 2002, p.2)	4C. *Limited Figural Drawing* (Bennett, 1993, p.44)	5C. *Cloze-Procedure* (Osterlind, 1998, p.242)	6C. *Concept Map* (Shavelson, R. J., 2001; Chung & Baker, 1997)	7C. *Discussion, Interview* (Bennett, 1993, p.45)
1D. *Multiple Choice with New Media Distractors* (Parshall et al, 2002, p.87)	2D. *Complex Multiple Choice* (Haladyna, 1994c, p.57)	3D. *Assembling Proof* (Bennett, 1993, p.44)	4D. *Bug/Fault Correction* (Bennett, 1993, p.44)	5D. *Matrix Completion* (Embretson, S, 2002, p. 225)	6D. *Essay* (Page et al, 1995, 561-565) *&* *Automated Editing* (Breland et al, 2001, pp.1-64)	7D. *Diagnosis, Teaching* (Bennett, 1993, p.4)

Cautions about media inclusion include: "Any new feature that is added to a test that is not essential to the variable the test is intended to measure is a potential threat to validity" (Van der Linden, 2002, p. 94). Others remind assessment developers that items should not require the examinee "to spend time producing irrelevant data (from the perspective of the measurement goal) or doing irrelevant mental processing" (Stout, 2002, p. 109). Stout uses the example of a five-minute video clip as a prompt for a five-minute creative essay as perhaps excessive. In answer to this, however, it seems relevant to remember that construct validity is only one goal of assessment. Other goals may include a better match with classroom activities and objectives, need for more authentic stimuli and responses (and that are engaging as classroom activities), and other aims that may be better satisfied by such media inclusion.

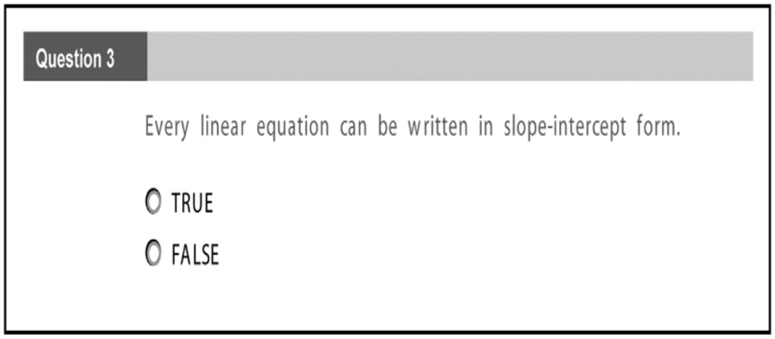
Additional innovations concerning the "observation" space have to do with automatic development of items. Automatic item generation is outside the scope of this paper, but it involves approaches such as templates for items or "item shells" used to create many items from one (Haladyna, 1994a) or in what we call item modeling, in which descriptions of task aspects, or facets, are devised and then implemented across a variety of contexts, or scenarios, such as can be modeled with some psychometric approaches (Rijmen & De Boeck, 2002).

The sections below briefly present characteristics and literature regarding each category of constraint in the Taxonomy table. We call the series of item types shown as "iconic" in that they represent categories of types which themselves can be modified to generate numerous useful designs. For the sake of brevity, we limit the iconic types shown here to four per each constraint category, and we provide some examples based on our work in mathematics and science assessment. References cited in the Taxonomy table include additional examples in other disciplines.

"Observation" Innovations in Category 1: Multiple Choice

Items that require an examinee to choose an answer from a small set of response options fall into the first column of the Taxonomy table, which is the multiple choice category (Bennett, 1993). Examples of four iconic types in this category are shown in Figures 2 to 5. These include the simplest selected response item types that offer only two choices, such as simple true/false items or Types 1A and 1B in the *Intermediate Constraint Taxonomy* presented in Table 1 on page 9 (Haladyna, 1994c). In the Type 1A example, respondents are asked whether it is true or false that a linear mathematical equation can be written in slope and intercept form, or in other words, as $y = mx + b$ (Figure 2). The correct answer in this case is *true*. Making a selection between "yes" and "no" for a given statement is one of the simplest and most constrained selected choice formats.

Figure 2: Example of Multiple Choice, item Type 1A: *True/False*[5]



Question 3

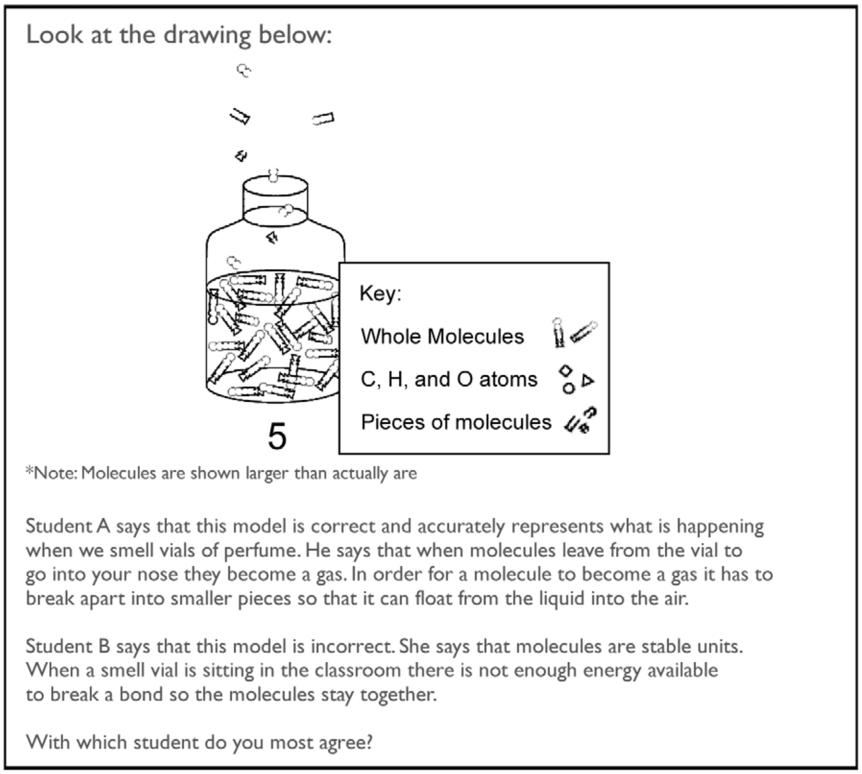
Every linear equation can be written in slope-intercept form.

○ TRUE

○ FALSE

Alternate choice items are similar to true/false items; however, rather than asking whether a single statement is correct or not, alternate choice offers two statements and asks the respondent to select the better option. Choices are often scenarios or cases, as shown in the Type 1B example in Figure 3 below. Here, students are shown two possible models for evaporation and must choose the most accurate response option. In this case, the correct answer is *Scenario B*.

Figure 3: **Example of Multiple Choice, item Type 1B: *Alternate Choice*[6]**



Look at the drawing below:

5
Key:
Whole Molecules
C, H, and O atoms
Pieces of molecules

*Note: Molecules are shown larger than actually are

Student A says that this model is correct and accurately represents what is happening when we smell vials of perfume. He says that when molecules leave from the vial to go into your nose they become a gas. In order for a molecule to become a gas it has to break apart into smaller pieces so that it can float from the liquid into the air.

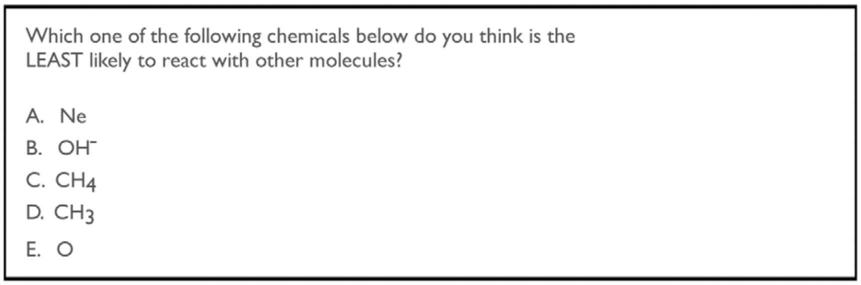
Student B says that this model is incorrect. She says that molecules are stable units. When a smell vial is sitting in the classroom there is not enough energy available to break a bond so the molecules stay together.

With which student do you most agree?

True/false and alternate choice formats can have limited utility as assessment formats because of the high incidence of guessing (Haladyna, 1994c). Guessing can be handled in different ways, including offering a series of true/false questions and then scoring over the group (Type 2A in Table 1 on page 9). Alternatively, some instruments attempt to handle guessing with statistical adjustments.

As the available choices from which to select answers increase beyond two, Type 1C items are generated, which are the conventional or standard multiple choice questions with usually four or five distractors and a single correct option. The example presented in Figure 4 below shows a list of chemicals and asks which chemical is likely to be least reactive. The answer requires understanding valence electron structure and that the least reactive is neon *(Ne)*.

Figure 4: **Example of Multiple Choice, item Type 1C: *Conventional Multiple Choice*[7]**



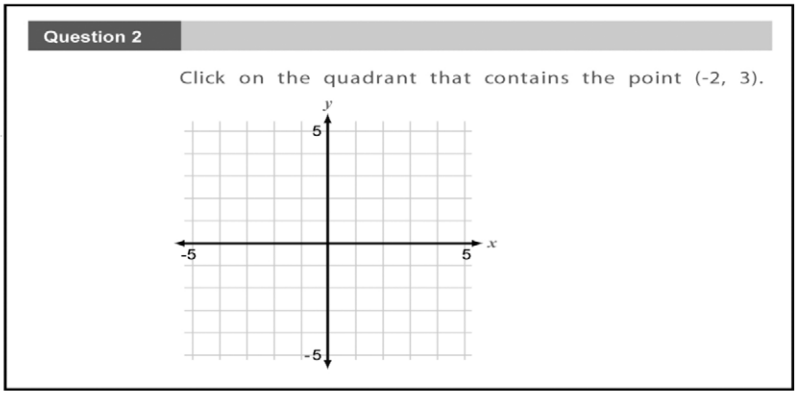
Which one of the following chemicals below do you think is the LEAST likely to react with other molecules?

A. Ne
B. OH^-
C. CH_4
D. CH_3
E. O

Strengths and weaknesses of the standard multiple choice item were outlined in the introduction to this paper. Multiple choice questions are readily scorable and less susceptible to guessing in comparison to true/false and alternate choice formats since more answer choices are available. However, overuse or exclusive use of multiple choice has been criticized as decontextualizing and decomposing learning, encouraging poor habits of mind and reducing the richness of the instructional experience by teaching to selection-only test formats (Resnick & Resnick, 1992). Additionally, we believe limiting computer-based approaches to standard multiple choice items may not optimize the assessment potential of the technology platform, where more construction can be readily achieved.

Innovations in the multiple-choice category for online settings can include new response actions not common in paper-and-pencil settings, such as clicking on an area of a graphical image, and can also include new media, such as distractors that are sound clips (Parshall, Spray, Kalohn, & Davey, 2002). Such new media innovations are represented in our Taxonomy as Type 1D, Multiple Choice with New Media Distractors, and an example is given of a point-and-click multiple choice item (Figure 5). In this example, respondents must select one of the four quadrants shown on an *x-y* coordinate plane. There are four quadrants from which to choose, so this is analogous to the standard multiple choice question with four possible responses and one correct choice, but with the mode of response involving graphical action.

Figure 5: Example of Multiple Choice, item Type 1D: *Multiple Choice with New Media Distractors*[8]



Question 2

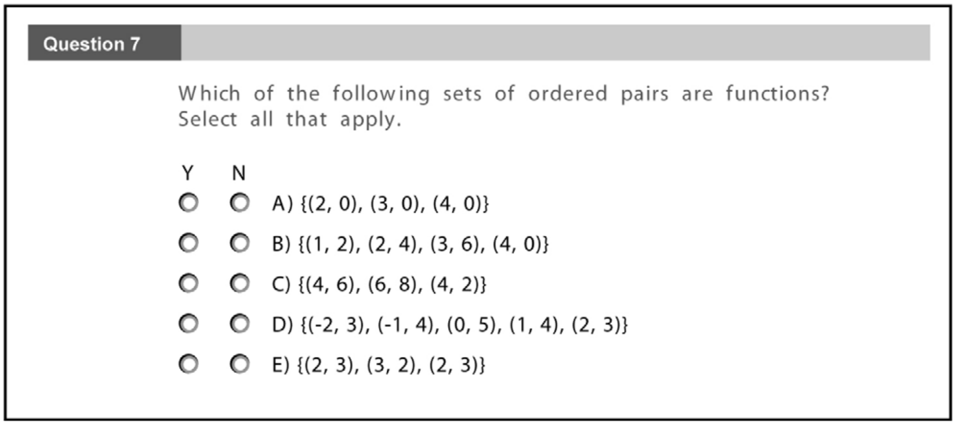
Click on the quadrant that contains the point (-2, 3).

"Observation" Innovations in Category 2: Selection/Identification

As the list of distractors or choices in a multiple-choice item grows and becomes rich enough to limit drastically the chances of guessing a correct answer, item types can be classified as selection/identification approaches (Figures 6 to 9).

Type 2A, multiple true-false (MTF), is really an item set, or item bundle, that offers the advantage of administering many items in a short period of time, but with a single score over many items so that guessing is controlled within the item group. It is unlikely for a respondent to randomly guess consistently correct over a set of items. The example given in Figure 6 lists sets of numbers and asks which sets could be generated by a given function. In this example, the key to a successful answer is understanding that for a function, one input number should give only one output number. Thus for each unique value of *x*, there should be only one possible *y*. This rules out answers C and E, leaving A, B and D as the true statements to select.

Figure 6: Example of Selection/Identification, item Type 2A: *Multiple True/False*[9]



Question 7

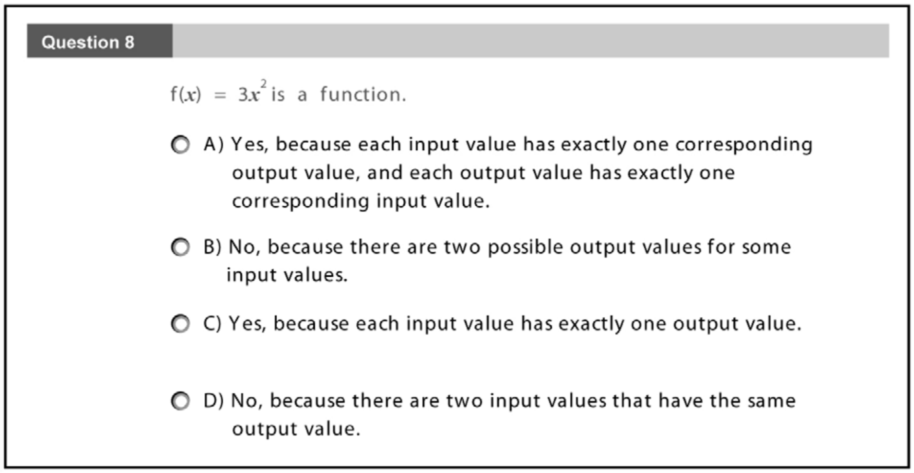
Which of the following sets of ordered pairs are functions? Select all that apply.

Y N
○ ○ A) {(2, 0), (3, 0), (4, 0)}
○ ○ B) {(1, 2), (2, 4), (3, 6), (4, 0)}
○ ○ C) {(4, 6), (6, 8), (4, 2)}
○ ○ D) {(-2, 3), (-1, 4), (0, 5), (1, 4), (2, 3)}
○ ○ E) {(2, 3), (3, 2), (2, 3)}

Research on the MTF format generally supports its use, except for the detriment of unfamiliarity by item writers (Haladyna, 1994c). According to Haladyna (1994c), MTF and conventional multiple choice (MC) were compared in a medical testing situation that found MTF yielded more reliable scores. However, conventional multiple-choice was more highly correlated with complex measures of competence and, at least in the examples studied, MTF items seemed to be a measure of more basic knowledge.

The yes/no explanation format, Type 2B, goes one step beyond being a set or bundle of true/false questions. It requires a two-step reasoning process involving identifying which alternative is correct and then recognizing why it is correct, for each true/false pair (McDonald, 2002). The example shown in Figure 7 makes a mathematical statement and asks (i) whether or not it is correct and (ii) why it is or is not correct. The correct answer is *C*.

Figure 7: **Example of Selection/Identification, item Type 2B: *Yes/No with Explanation*[10]**



Question 8

$f(x) = 3x^2$ is a function.

○ A) Yes, because each input value has exactly one corresponding output value, and each output value has exactly one corresponding input value.

○ B) No, because there are two possible output values for some input values.

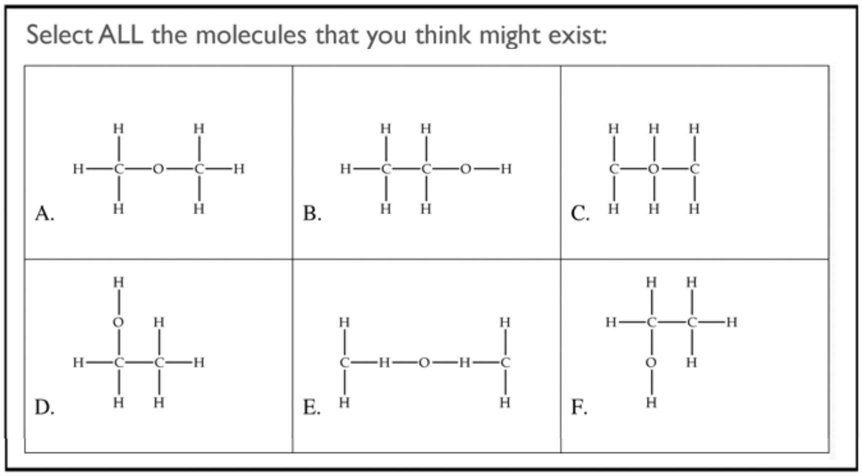
○ C) Yes, because each input value has exactly one output value.

○ D) No, because there are two input values that have the same output value.

McDonald (2002) cites this item type, if well-written, as tending to have higher discrimination indices, or more readily distinguishing between students of higher and lower abilities, than conventional multiple choice. It is challenging to write well, though, as each set of explanations must be plausible for each pair choice. Otherwise we believe students may have clues as to which answer is correct just based on logical reasoning and elimination of unlikely explanations.

Type 2C in the selection/identification category is the multiple answer or *multiple mark* format, which includes, for example, a medical exam item that prompts examinees to select all of the elements listed that are symptoms of a particular disease (Parshall, Spray, Kalohn, & Davey, 2002). The example shown in Figure 8 involves selecting viable molecules from among a set of possible structures.

Figure 8: Example of Selection/Identification, item Type 2C: *Multiple Answer*[11]

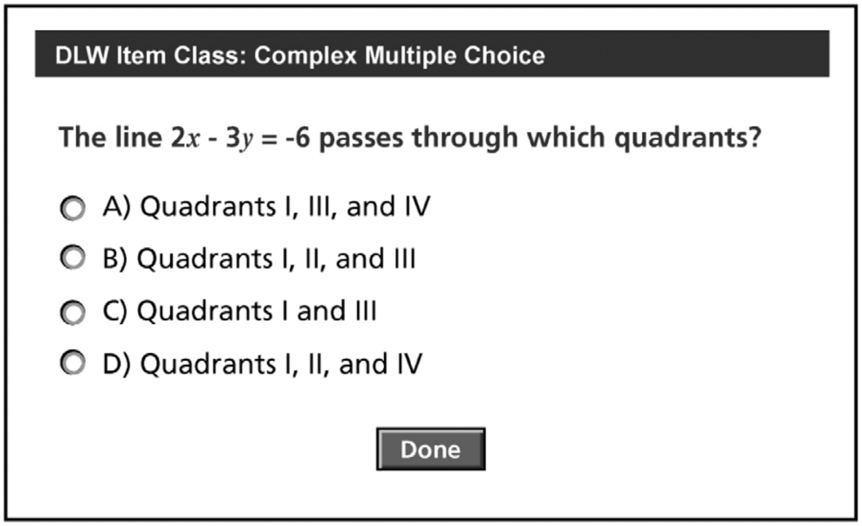


Select ALL the molecules that you think might exist:

According to Haladyna (1994c), the multiple mark format has a historical record of use but has been neglected in recent decades in formal testing settings. He said that students who guess on this format tend to make errors of omission, which may introduce bias in test scores and needs more research to understand as a test effect. He also concluded that patterns of response, desirable number of distractors and other aspects are not well understood for this item type. Still, multiple answer "gets good grades in terms of performance when compared with other multiple-choice formats," according to Haladyna (1994c, p. 60). He predicts this item type will come to be used more in both classroom testing and formal testing programs and finds the main attraction is presenting many items in a short period of time. We believe an additional attraction in e-learning contexts is that missing multiple marks can be tracked and queried with adaptive prompts.

The final type shown in this category is Type 2D, complex multiple choice, in which combinations of correct answers are offered as distractors. The example shown in Figure 9 involves determining through which quadrants a particular line passes on an *x-y* coordinate plane.

Figure 9: Example of Selection/Identification, item Type 2D: *Complex Multiple Choice*



DLW Item Class: Complex Multiple Choice

The line $2x - 3y = -6$ passes through which quadrants?

○ A) Quadrants I, III, and IV

○ B) Quadrants I, II, and III

○ C) Quadrants I and III

○ D) Quadrants I, II, and IV

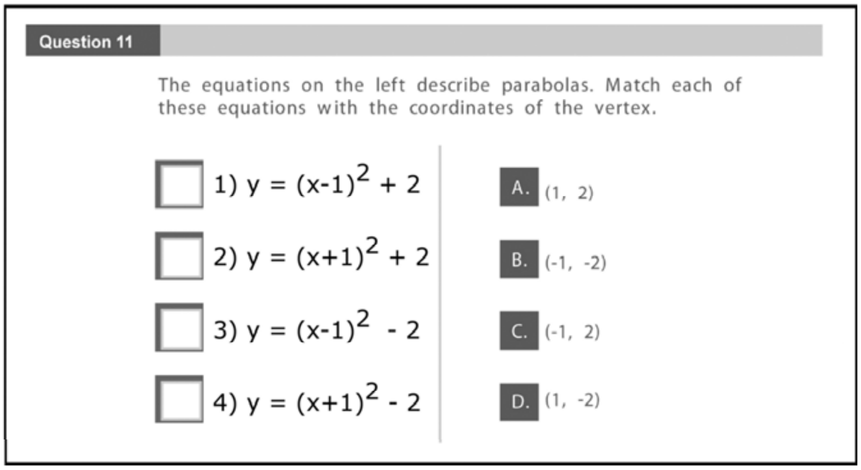
This item type was first introduced by the Educational Testing Service (ETS) and became popular in some formal testing programs (Haladyna, 1994c). However, Haladyna says several studies have found it to be an inferior item type because, among other reasons, examinees with better test-taking skills leverage knowledge of one option as absolutely correct or incorrect to eliminate distractors and improve their guessing ability.

"Observation" Innovations in Category 3: Reordering/Rearrangement

Once again in this category, responses are chosen from a supplied set, but this time with the additional construction requirement for the respondent of rearranging the choices into an order or sequence requested by the item stimulus (Figures 10 to 13). Given the richness of media inclusion and possible new response actions in computer environments, sequencing and ranking are becoming popular in courseware activities in computer environments. However, in the 44 papers reviewed to contribute to this Taxonomy, there were few references to ranking and sequencing as formal item types, and no discussions of potential merits and drawbacks from psychometric perspectives.

We show four iconic types of sequencing/rearrangement designs in our Taxonomy. Type 3A, Figure 10, involves simple pair matching of item stems on the left of the screen with a set of possible responses on the right. This matching item type is a popular format in classroom-based assessment but rare in large-scale testing programs (Haladyna, 1994c).

Figure 10: Example of Reordering/Rearrangement, item Type 3A: *Matching*[12]



Question 11

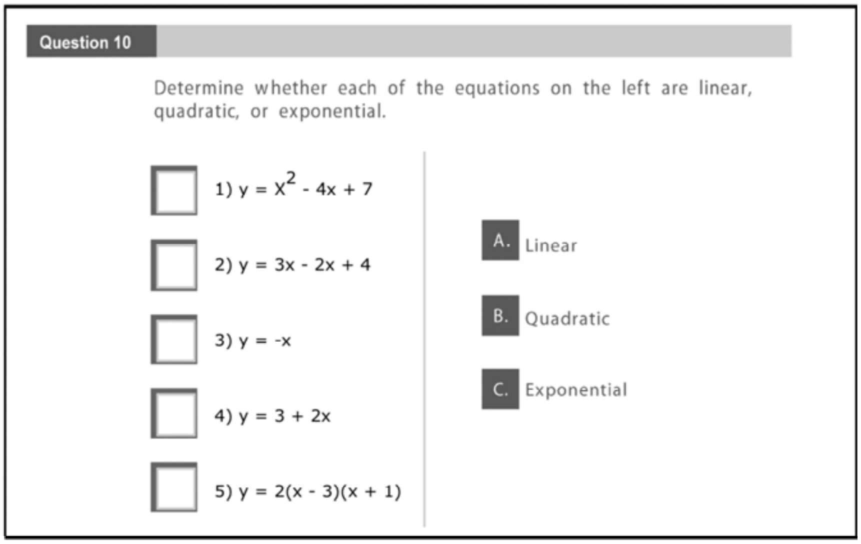
The equations on the left describe parabolas. Match each of these equations with the coordinates of the vertex.

□ 1) $y = (x-1)^2 + 2$	A. (1, 2)
□ 2) $y = (x+1)^2 + 2$	B. (-1, -2)
□ 3) $y = (x-1)^2 - 2$	C. (-1, 2)
□ 4) $y = (x+1)^2 - 2$	D. (1, -2)

Haladyna (1994c) says there is little research to report on unique features for the matching item format. He recommends such items continue to be used as a variation of conventional multiple-choice since they are easy to construct and administer; they lend themselves to testing associations, definitions and examples; and they are efficient in space, as options do not have to be repeated, and in test-taker time. Besides the lack of research and theory, other limitations for the matching type come with item-writing traps that are easy to fall into, including nonhomogeneous options, such as mixing sets of things, people and places, or providing equal numbers of items and options, both of which make guessing easier and can bring test-taking skills into play as a nuisance, or unwanted, dimension of performance.

Type 3B involves categorizing an object, concept or other construct into an appropriate parent or umbrella class, such as categorizing elements in a list. The example shown in Figure 11 requires the categorization of a particular instance of a mathematical equation into families of equations: linear, quadratic or exponential.

Figure 11: Example of Reordering/Rearrangement, item Type 3B: *Categorizing*[13]



Question 10

Determine whether each of the equations on the left are linear, quadratic, or exponential.

□ 1) $y = X^2 - 4x + 7$

□ 2) $y = 3x - 2x + 4$

□ 3) $y = -x$

□ 4) $y = 3 + 2x$

□ 5) $y = 2(x - 3)(x + 1)$

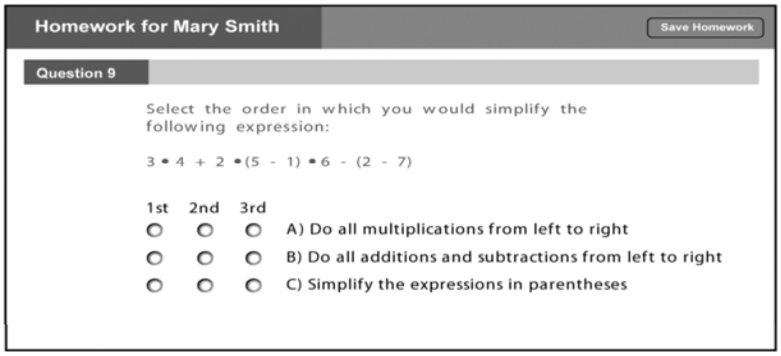
A. Linear

B. Quadratic

C. Exponential

Type 3C involves indicating the order or sequence in which a series should occur. Bennett (1993) mentions ordering a list of words into a logical sentence and arranging a series of pictures in sequence as examples of reordering/rearrangement. The example we provide in Figure 12 requires the respondent to sequence the order of operations to solve a mathematics problem.

Figure 12: Example of Reordering/Rearrangement, item Type 3C: *Ranking & Sequencing*[14]



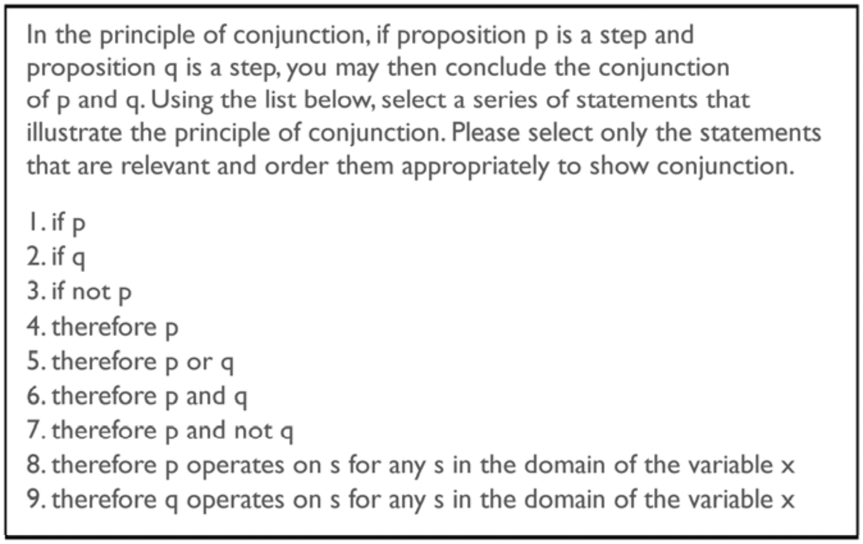
Select the order in which you would simplify the following expression:

3 • 4 + 2 •(5 - 1) •6 - (2 - 7)

1st	2nd	3rd	
○	○	○	A) Do all multiplications from left to right
○	○	○	B) Do all additions and subtractions from left to right
○	○	○	C) Simplify the expressions in parentheses

Type 3D also involves arranging a series, but in this case the arrangement itself forms an explanation, or proof. This is most commonly used in assembling mathematical proofs, or in other words using mathematical evidence to prove a statement of interest, given identified premises and a list of axioms from which to choose a series for the proof. A very simple example of a proof that can be assembled is shown in Figure 13, but proofs and axiom choices can become much more complex in real items. Here the correct choices for the sequence of the proof are *(i) if p, (ii) if q, (iii) therefore p and q*. The other choices that are offered are not needed for the proof.

Figure 13: Example of Reordering/Rearrangement, item Type 3D: *Assembling Proof* [15]



In the principle of conjunction, if proposition p is a step and proposition q is a step, you may then conclude the conjunction of p and q. Using the list below, select a series of statements that illustrate the principle of conjunction. Please select only the statements that are relevant and order them appropriately to show conjunction.

1. if p
2. if q
3. if not p
4. therefore p
5. therefore p or q
6. therefore p and q
7. therefore p and not q
8. therefore p operates on s for any s in the domain of the variable x
9. therefore q operates on s for any s in the domain of the variable x

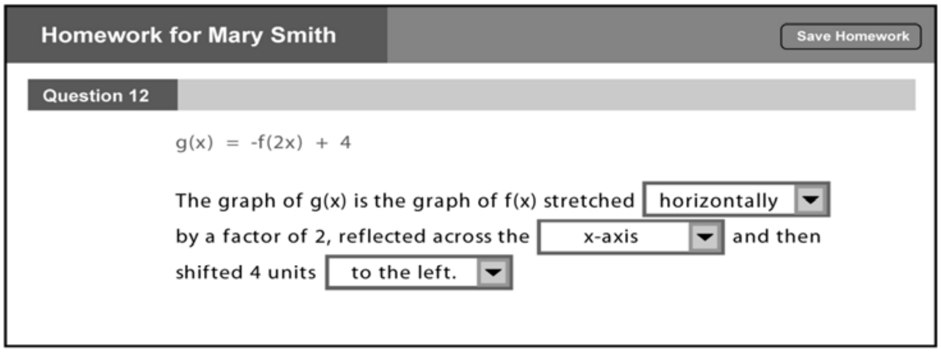
In general, reordering and rearrangement items in all these Type 3 categories can be scored as dichotomous – correct if an ideal sequence was achieved and incorrect otherwise – or it might be possible to award partial credit based on permutations of the sequence, where some incorrect permutations are valued and scored more highly than others. Some intriguing work in cognition may involve whether certain patterns of incorrect answers can diagnose patterns of reasoning (Scalise, 2004). Theory, scoring rubrics or empirical research would need to justify making such valuing distinctions in the score permutations of incorrect sequences.

“Observation” Innovations in Category 4: Substitution/Correction

Category 4 items, involving substitution and correction tasks, require an additional degree of construction by the respondent. In Categories 1 to 3 item types, the respondent knew one or more of the provided answers would be necessary to correctly respond to the item. In substitution and correction, the item may or may not need to be changed in each location identified. Therefore the respondent must not only identify the correct answer to select, but also whether any of the provided solutions should be used at all.

Substitution and correction tasks come in a variety of formats. Item Type 4A, the interlinear format, offers substitution choices interlinearly, or within each line of the item. Typically one choice is to leave the line unchanged. The interlinear format is shown in Figure 14, in which drop down menus offer the word choices embedded within lines. The example involves changing word choice in explaining how a line is different between two formulas. Originally the format was created for measuring writing skills via a multiple-choice format that was efficient to score, but, according to Haladyna, it was determined not to have high fidelity (Haladyna, 1994c). That is, the multiple-choice version they used lacked construct validity for measuring the skill of writing it was intended to assess.

Figure 14: Example of Substitution/Correction, item Type 4A: *Interlinear*[16,17]

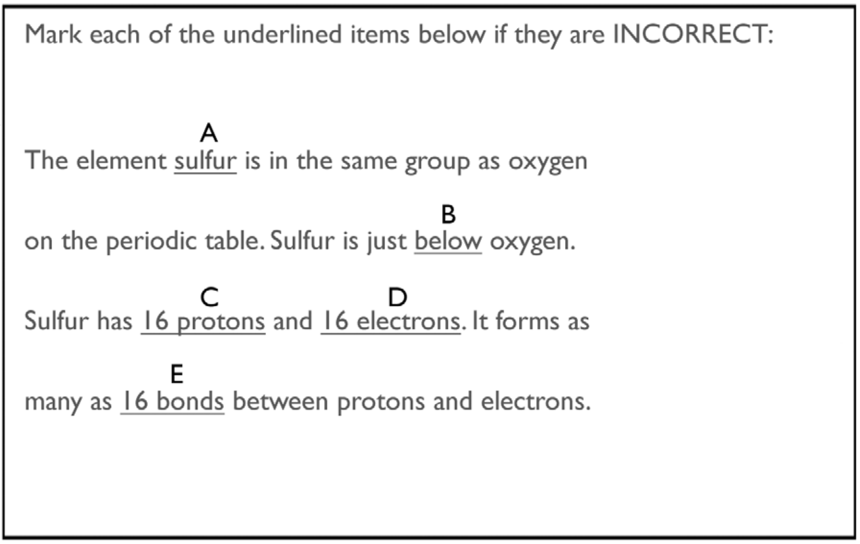


g(x) = -f(2x) + 4

The graph of g(x) is the graph of f(x) stretched horizontally by a factor of 2, reflected across the x-axis and then shifted 4 units to the left.

Type 4B, the sore finger exercise, is similar but each item choice for correction stands alone. In the example given in Figure 15, regarding characteristics of the element sulfur, statements A–D are correct but E is incorrect. The student should select *E*.

Figure 15: Example of Substitution/Correction, item Type 4B: *Sore-finger*[18]



Mark each of the underlined items below if they are INCORRECT:

A

The element sulfur is in the same group as oxygen

B

on the periodic table. Sulfur is just below oxygen.

C D

Sulfur has 16 protons and 16 electrons. It forms as

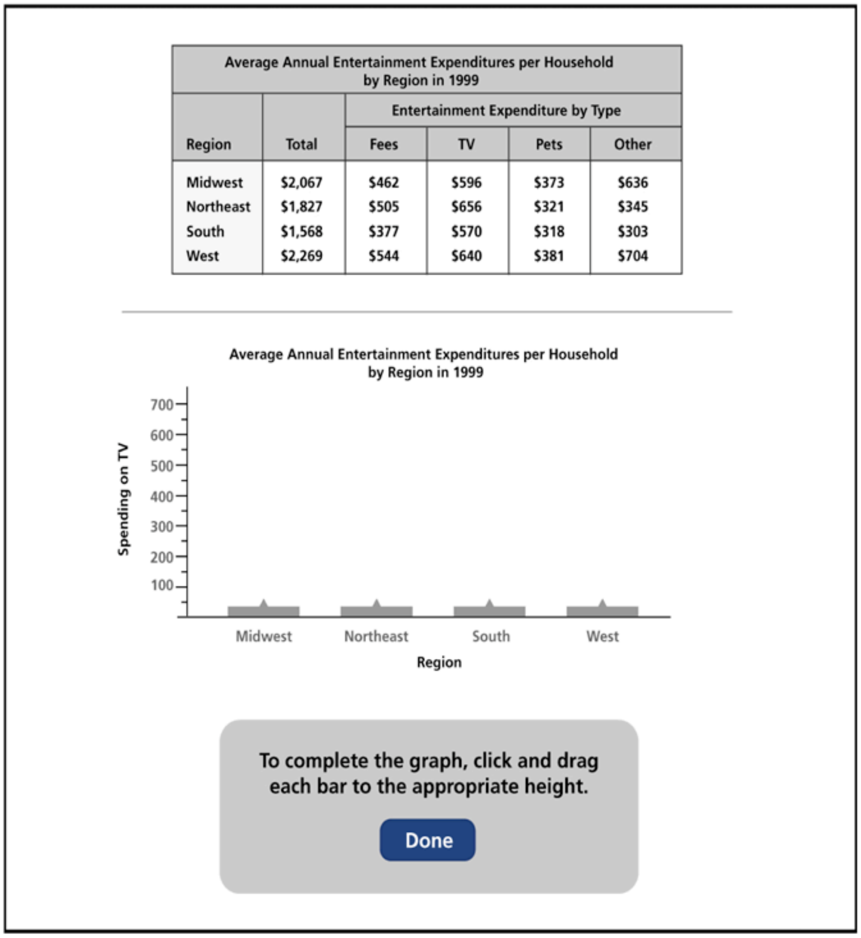
E

many as 16 bonds between protons and electrons.

Note: INCORRECT answers are selected on screen by point and click.

In type 4C, limited figural drawing, a small part of a drawing, illustration or other graphic element is adjusted or corrected (Bennett, Sebrechts, & Yamamoto, 1991). The example given in Figure 16 shows four columns of a bar graph in place on a figure, but the bars are not adjusted to the right heights for the data given. The student must adjust them.

Figure 16: Example of Substitution/Correction, item Type 4C: *Limited Figural Drawing*[19]



Average Annual Entertainment Expenditures per Household by Region in 1999

		Entertainment Expenditure by Type			
Region	Total	Fees	TV	Pets	Other
Midwest	$2,067	$462	$596	$373	$636
Northeast	$1,827	$505	$656	$321	$345
South	$1,568	$377	$570	$318	$303
West	$2,269	$544	$640	$381	$704

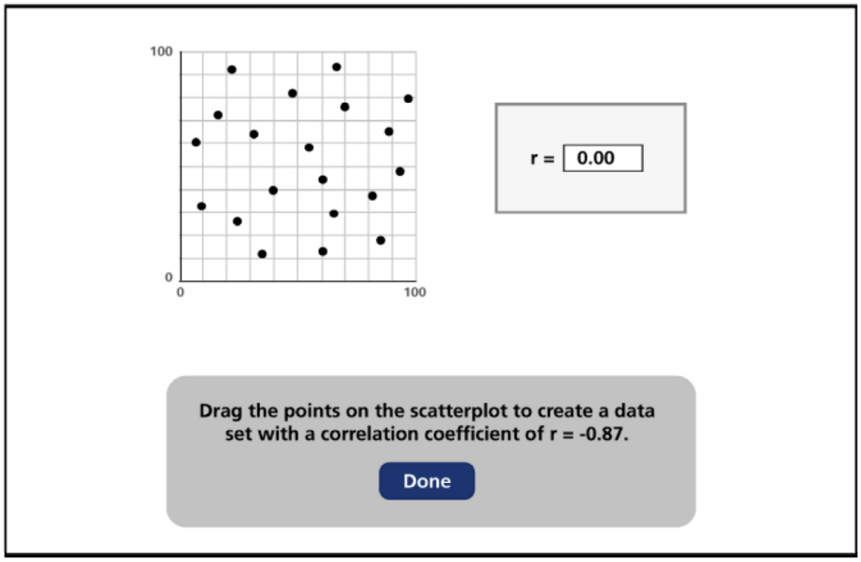
Average Annual Entertainment Expenditures per Household by Region in 1999

To complete the graph, click and drag each bar to the appropriate height.

Limited figural drawing can be seen as a graphical analog to the substitution or correction of a word or math problem, and has become of great interest in online education products and courseware as it can take advantage of rich media inclusion but remain automatically scorable. Little formal research on the psychometric properties of this type is yet available.

Type 4D, bug or fault correction, requires that some part of a problem solution be changed to bring the answer into compliance with a condition that has been set for the task. The example shown in Figure 17 below involves reconfiguring a pattern of data to change a summary statistic regarding the data. Corrections of simple bugs in algorithmic schemes for computer programming code and other editing exercises, such as changing diagramming for network wiring and circuitry, are common in the bug/fault correction format.

Figure 17: Example of Substitution/Correction, item Type 4D: *Bug/Fault Correction*[20]



Drag the points on the scatterplot to create a data set with a correlation coefficient of r = -0.87.

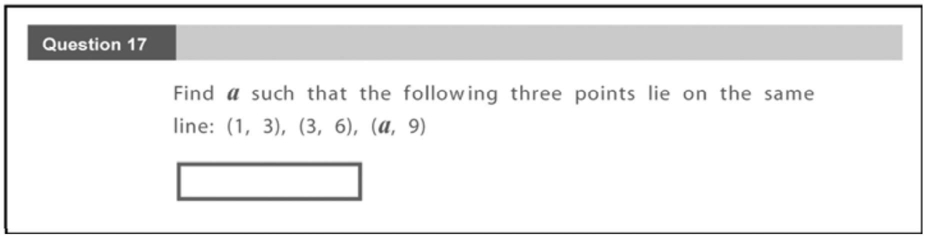
"Observation" Innovations in Category 5: Completion

The remaining categories of the constraint taxonomy involve items in which not all the components for the answer are supplied in advance, or in which the entire problem-space is contained in the set of answers. The assessment literature tends to treat all these formats as innovative types, or alternative assessments, to a greater or lesser degree.

The completion category asks respondents to finish an incomplete stimulus, Figure 6 on page 15. Item types include single numerical constructed items, short-answer and sentence completion, Cloze-procedure, and matrix completion problems (Embretson, 2002). Much recent work on the completion format focuses on techniques of automatic scoring, which is not treated in this paper. A good treatment of this topic is available in a book on automated scoring, which includes a review chapter across numerous approaches (Scalise & Wilson, 2006).

Type 5A is the single numerical constructed item type, which asks examinees to calculate and supply a desired number. The example given in Figure 18 shows the coordinates of three points of a line, with one coordinate missing. The student must supply the missing number. Here issues of how far to extend the decimal point, for instance, can easily be accommodated in scoring algorithms.

Figure 18: Example of Completion, item Type 5A: *Single Numerical Constructed* [21]



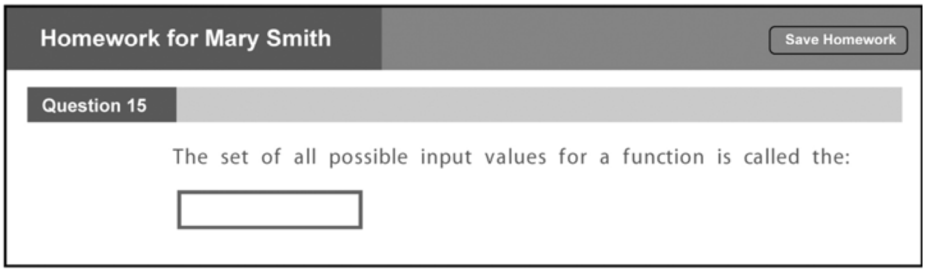
Question 17

Find *a* such that the following three points lie on the same line: (1, 3), (3, 6), (*a*, 9)

This item format was once assumed in the literature to be best for low task complexity (Parshall, 2002), but to us it seems perhaps an unnecessary limitation as items demanding complex problem-solving, strategy selection and solution construction can result in single, well-defined numerical answers. This is how the item type is often used in the classroom, although often with the specification that students show their work so that the problem-solving process is more clearly elucidated for partial credit scoring and learning intervention, and to discourage guessing without problem solving.

Type 5B, short-answer and sentence completion, is sometimes called the fill-in-the-blank format, though that term is avoided in the literature. The example given in Figure 19 asks the student for the name of a particular set of numbers. The correct answer is "domain." The format has traditionally been considered to test mainly factual recall, as the respondent is only allowed to supply a word or short phrase (Osterlind, 1998). However, it seems reasonable, although not discussed in the literature reviewed, that computer-based approaches perhaps can allow for more scoring options, or in other words an expanded outcome space, since an extensive databank of acceptable responses can be built, perhaps allowing for richer use of the item.

Figure 19: Example of Completion, item Type 5B: *Short-Answer & Sentence Completion*[22]



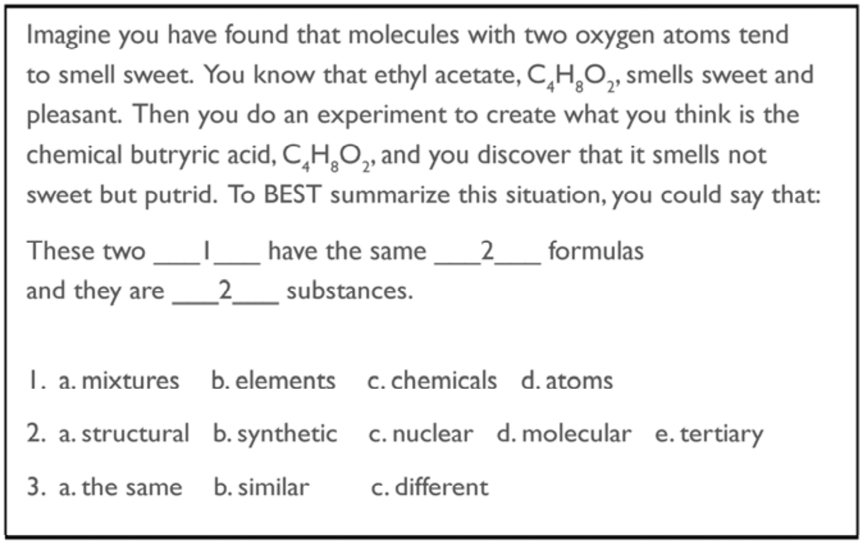
The set of all possible input values for a function is called the:

Osterlind (1998) says short answer items are presumed to reduce guessing, but there is little research to support this point. He suggests caution in the assumption as "examinees rarely guess blindly on an item" (p. 239) and many items "logically lead an examiner to one – and just one – correct response" (p. 239).

Item writing can be a big challenge in this type, he says. Not only can the outcome space be too narrowly constructed, so as to allow for high guessing rates as above, but it also can be too widely conceived so that the student answer is correct but remains quite off the topic from what was expected, or what was being measured. This is where computer-based approaches that attempt to capture and categorize or analyze a range of empirical responses may make the item type more valuable.

Type 5C, Cloze-procedure, by contrast, is "certainly one of the most widely researched item types" (Osterlind, 1998, p. 250). It assesses, for instance, reading ability by embedding test items within a passage. Words from the passage are deleted and students supply answers by selecting words from a list of alternatives, thus the passage looks like fill-in-the-blank but the response alternatives are multiple choice. The example given in Figure 20 shows Cloze in a science context, with three deletions in the passage and several choices for each deletion (more might be offered to expand the outcome space). The example contrasts two different molecules. First some information is supplied about the molecules, then the Cloze passage removes three key words that contrast the molecules. The correct answers would be *1c (chemicals), 2d (molecular) and 3a (different)*. The key knowledge for students to construct here is that the molecules contain the same atoms, thus having the same molecular formulas, but the atoms are arranged in different patterns relative to each other, generating different chemicals with different smell properties.

Figure 20: Example of Completion, item Type 5C: *Cloze-procedure*[23]



Imagine you have found that molecules with two oxygen atoms tend to smell sweet. You know that ethyl acetate, $C_4H_8O_2$, smells sweet and pleasant. Then you do an experiment to create what you think is the chemical butyric acid, $C_4H_8O_2$, and you discover that it smells not sweet but putrid. To BEST summarize this situation, you could say that:

These two ____1____ have the same ____2____ formulas
and they are ____2____ substances.

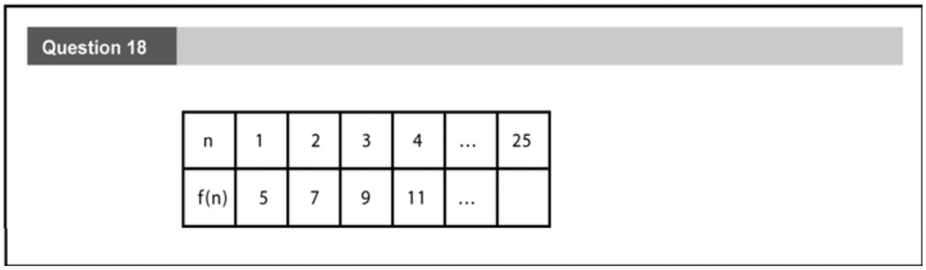
1. a. mixtures b. elements c. chemicals d. atoms
2. a. structural b. synthetic c. nuclear d. molecular e. tertiary
3. a. the same b. similar c. different

Cloze was developed to straddle the need for longer and more realistic reading passages in an item format that can still be readily scored. Once again, Osterlind emphasizes the difficulty of writing good items in this format. Passages must contain all the information examinees need to respond correctly to embedded items yet all the alternate word choices must be plausible within the passage (Osterlind, 1998).

There is no set rule for word removal in Cloze. One typical approach is to delete every fifth word when reading itself is being assessed. Then the selected words usually are common and their meaning easily recognizable so that the response is a test of reading comprehension rather than specific understanding of individual vocabulary words available in the distractors. In other areas, key words of interest are often deleted rather than words that fall at a specific deletion frequency. In either case, all distractors should work in the sentence grammatically if read in isolation, but only one should be appropriate when read in the context of the complete passage. The *1986 College Entrance Examination Board DRP Handbook* claims good validity for the type when well written, as "choosing wrong answers signals a failure to comprehend" (Osterlind, 1998, p. 250).

Type 5D, the matrix completion format, present a matrix of patterns with one or more cells left blank. Respondents are asked to fill the empty cells from a set of supplied answers. Matrix completion has an extensive history in intelligence measurement and has been used in various tests of pattern recognition, correspondence and generation (Embretson, 2002). In the example given in Figure 21 below, respondents must determine the function that generates a series of numbers and then use the function to complete the missing cell in the series. The cells containing ellipses (...) indicate that the series continues until the given series element is reached. Since the equation here can be determined to be $y = 2x + 3$, then when x equals 25 the missing value of y is 53, which should be input to complete the matrix pattern.

Figure 21: Example of Completion, item Type 5D: *Matrix Completion*[24]



Question 18

n	1	2	3	4	...	25
f(n)	5	7	9	11	...	

In computer-environments, a somewhat different conception of matrix completion is proving to be a valuable item type. The matrix is a table or spreadsheet and correct patterns, which can easily be graphics, words or numbers, as well as sound clips, film clips and animations, are dragged to the appropriate empty cells, similar to formal matrix completion. But rather than more abstract patterns used in intelligence testing, the pattern selection in this usage assesses domain-rich knowledge, such as in a genetics courseware product supplying a genotype to match a phenotype or dragging a correct graphic to a certain cell to represent a pattern or principle. The item type allows for a great deal of flexibility in the task assignment, openness of response and media inclusion, but is readily computer-scorable, making it potentially a powerful item type in computer environments.

It can be seen that depending on what is called for in matrix completion, the matrix type can fall into a number of categories in the Taxonomy in Table 1 on page 9, for instance reordering, substitution and construction, as well as simple completion. Thus this type blurs the lines of the constraint-based item taxonomy. Domain-specific matrix completion tasks may be among the families of innovation most ripe for computer-based applications.

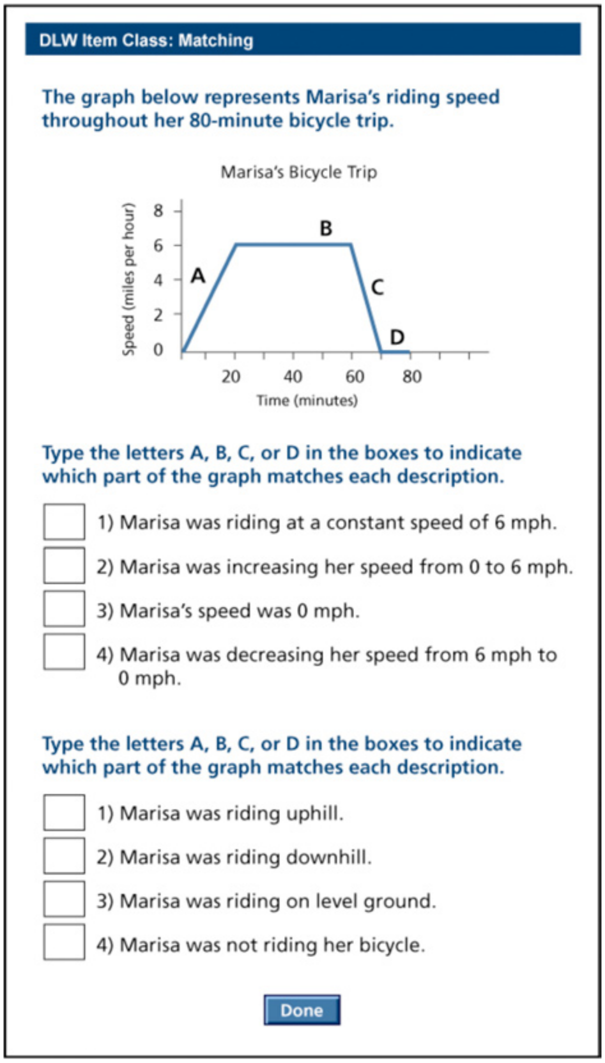
"Observation" Innovations in Category 6: Construction

As the level of constraint moves into the entirely constructed response, rating and scoring become increasingly greater challenges. ETS researchers define a complex constructed response as denoting "a task whose solutions are composed of many elements, take a variety of correct forms, and, when erroneous, approximate accurate answers to different degrees" (Bennett, Sebrechts, & Yamamoto, 1991, p. 1). A few simple examples are shown in Figures 22 to 25 below.

One caution in using constructed response is that sometimes these items are omitted more often by respondents. In the National Assessment of Educational Progress (NAEP), for instance, constructed-response tasks are omitted far more often than multiple-choice questions (Bennett, 1993).

The first item type listed in the construction category of the item Taxonomy, Type 6A, is open-ended multiple choice or uncued multiple-choice, which supplies a presumably uncued set of distractors (Haladyna, 1994b). The uncued aspect is somewhat unusual in the way it goes about creating opportunity for construction. Rather than having students originate and provide some portion of the answer to the question, selection choices are provided. However, the selection choices cover the entire problem space, which limits selection advantage or selection cues to the respondent and thus, arguably necessitates the complete construction of a solution. Consider the example of Type 6A in Figure 22.

Figure 22: Example of Construction, item Type 6A: *Open-Ended or "Uncued" Multiple Choice*[25]



DLW Item Class: Matching

The graph below represents Marisa's riding speed throughout her 80-minute bicycle trip.

Type the letters A, B, C, or D in the boxes to indicate which part of the graph matches each description.

□ 1) Marisa was riding at a constant speed of 6 mph.

□ 2) Marisa was increasing her speed from 0 to 6 mph.

□ 3) Marisa's speed was 0 mph.

□ 4) Marisa was decreasing her speed from 6 mph to 0 mph.

Type the letters A, B, C, or D in the boxes to indicate which part of the graph matches each description.

□ 1) Marisa was riding uphill.

□ 2) Marisa was riding downhill.

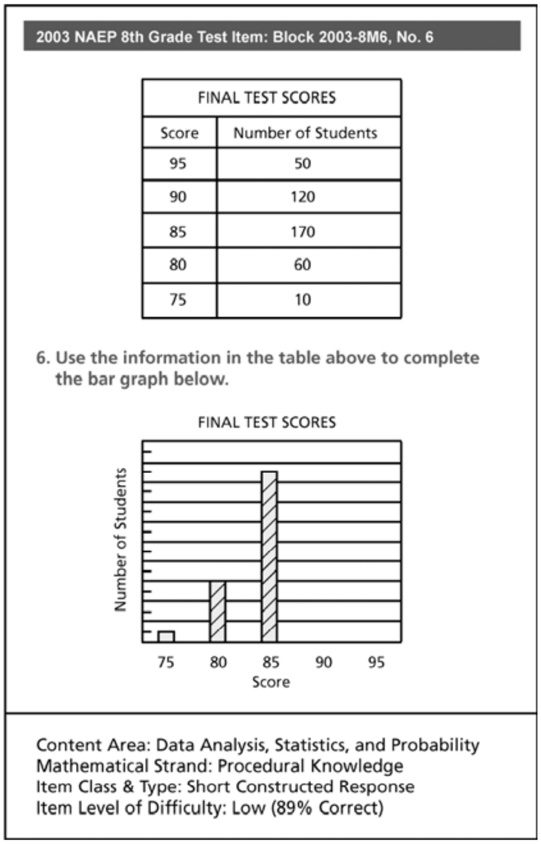
□ 3) Marisa was riding on level ground.

□ 4) Marisa was not riding her bicycle.

Here, students are presented with a line on a graph and asked to select the portion of the line related to each of several situations discussed. However, all of the possible line segments in the problem space are offered as possible answers. Students are not given any cues as to which segment is correct and thus must construct their answer from all possibilities. The answer to the first question is line segment B, but it could have been any other part of the line since all parts were offered in the selection choices[26].

Type 6B, figural constructed response, is a type of construction in which students draw in line segments or other graphic images. This is similar to the limited figural drawing of Type 4C, except limited figural drawing in Type 4C only required a substitution or correction to a representation while 6B allows for a fuller construction of the representation without presenting to the respondent the portion to be drawn in the figure given. The example shown in Figure 23 requires students to first, drag a bar onto a bar chart in the correct position along the *x*-axis and then drag the bar to the correct height along the *y*-axis.

Figure 23: Example of Construction, item Type 6B: *Figural Constructed Response*[27]



2003 NAEP 8th Grade Test Item: Block 2003-8M6, No. 6

FINAL TEST SCORES

Score	Number of Students
95	50
90	120
85	170
80	60
75	10

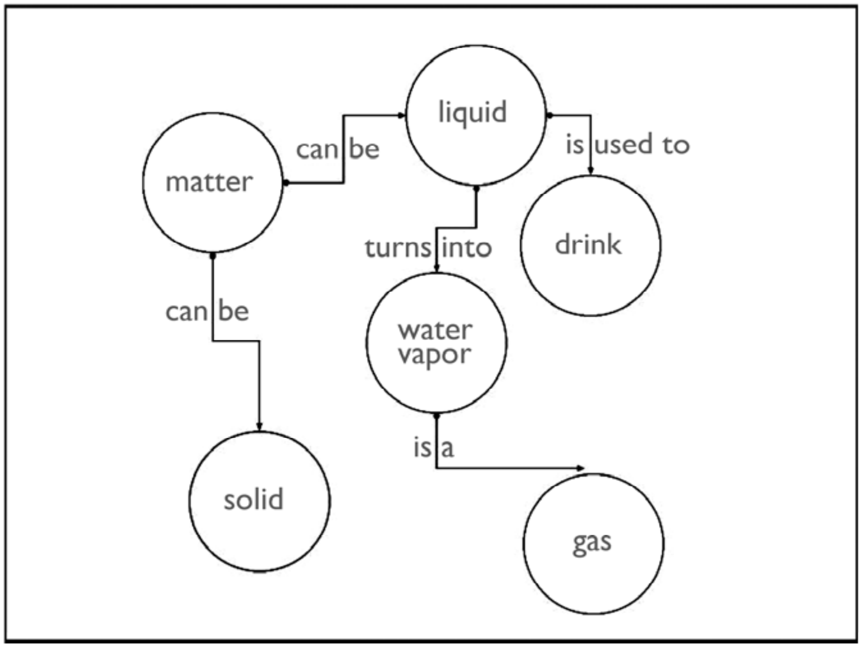
6. Use the information in the table above to complete the bar graph below.

FINAL TEST SCORES
Number of Students
75
80
85
90
95
Score

Content Area: Data Analysis, Statistics, and Probability
Mathematical Strand: Procedural Knowledge
Item Class & Type: Short Constructed Response
Item Level of Difficulty: Low (89% Correct)

Type 6C, the concept map, is also an interesting type of construction item to use in assessment (Shavelson, 2001). In Figure 24, students are given a list of terms and respond by arranging the terms relative to one another to indicate perceived relationships between the terms. The example given involves a respondent's conception of matter. Here, the respondent shows matter as having two phases or states: liquid and solid. Although this respondent does not recognize a third phase, gas, as connected to the matter concept directly, he or she does recognize that liquids, which are seen as matter, can become gases. This respondent also shows the category of liquid matter as generically being drinkable and becoming water vapor, suggesting that the respondent thinks of liquid phase matter primarily as water.

Figure 24: Example of Construction, item Type 6C: *Concept Map*[28]

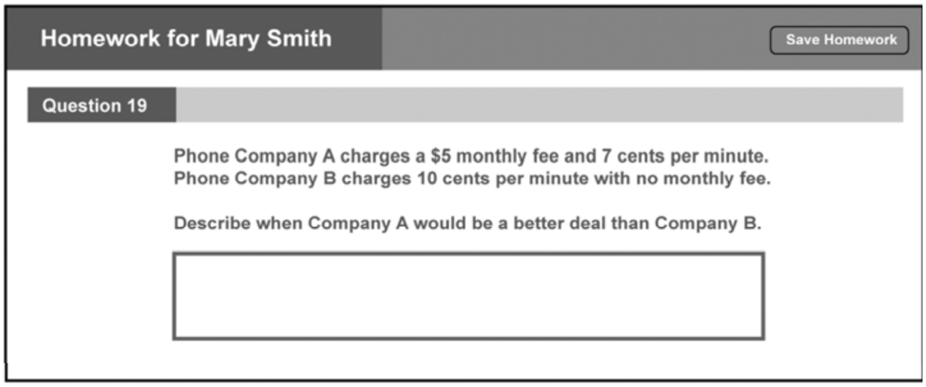


In concept maps, the construction component really comes in constructing relationships between and among concepts. The maps the students generate can be readily computer-scored or compared, yet arguably students have a great deal of opportunity to construct the map by arranging the terms in different configurations. The challenge here sometimes comes in interpreting the meaningfulness of differences in student maps and in how this information can be used in differentiating between students and their understanding of a topic.

Concept maps, a technique of latent semantics, are an entire field of active research in the area of cognition and are mentioned briefly here as a new way of measuring constructs distinct from more traditional, applied problem-solving approaches (Shavelson, Black, Wiliam, & Coffey, 2003). As items in assessment, concept maps investigate the relationships people hold between connected concepts. A more open construct would allow the inclusion of new topics to the set. Fixed-set items are scored on computer by noting and recording the relative placement of objects and links and by comparing these with maps generated by students at varying performance levels.

Type 6D, the essay item, is of course one of the most prevalent constructed response types used in assessment. The example given in Figure 25 shows a prompt that asks for an extended textual response or, in other words, a very short essay.

Figure 25: Example of Construction, item Type 6D: *Essay*[29,30]



Phone Company A charges a $5 monthly fee and 7 cents per minute.
Phone Company B charges 10 cents per minute with no monthly fee.

Describe when Company A would be a better deal than Company B.

Essay approaches online, of course, can vary widely (Page & Petersen, 1995). Most of the computer-based innovations in the essay format involve scoring, statistical models and rater analysis, for which a more thorough examination has been presented elsewhere (Scalise & Wilson, 2006). As examples, we will mention here two e-learning essay products: the ETS E-rater (Attali & Burstein, 2004) and C-rater products (Leacock, 2004). E-rater provides a holistic score for an essay, with a companion product, *Critique*, which provides near real-time feedback about grammar, mechanics and style as well as essay organization and development. E-rater employs a combination of feature analysis, vector space methods, and linguistic analysis to rate essays online. E-rater 2.0 was reviewed in a 2004 study that looked at scoring results from 6th to 12th grade users of *Criterion* as well from other examinations such as TOEFL (Test of English as a Foreign Language) essay data (Attali & Burstein, 2004). It reviewed students who completed two essays and compared the scores as test-retest scores. E-rater had a reliability of .60 as compared to .58 when human raters scored.

C-rater, by comparison, is intended to automatically analyze short-answer, open-ended response items. C-rater is designed to recognize responses that paraphrase correct responses, looking for syntactic variations, substitution of synonyms and the use of pronouns, as well as misspelling. It was reviewed in a 2004 study of a large-scale reading and math assessment of 11th graders in Indiana (Leacock, 2004). Over a sample of 170,000 short-answer responses in this study, C-rater was found to obtain 85 percent agreement with human raters.

Some constructed essay innovations attempt to measure language construction skills similar to essay revision rather than scoring of whole essays. GRE researchers have investigated automatic editing, which is a less constrained version of the old interlinear exercise discussed in category 4 under substitution and correction (Breland, Kukich, & Hemat, 2001). In automatic editing, an examinee is given a written passage with unspecified errors in grammar and sentence structure. The examinee's task is to identify and correct errors and write new words and phrases. Breland et al. (2001) describe the hope that replacement of one of two essays with automated editing could possibly raise reliabilities of about .70 to an estimated reliability of .84.

A big concern with automated essay scoring is whether the scoring criteria is a proxy to the true criterion of interest (e.g., average word length instead of word meaning), and thus coachable along alternate dimensions (Scalise & Wilson, 2006).

"Observation" Innovations in Category 7: Presentation/Portfolio

Falling into the least constrained or "presentation/portfolio" category of the item Taxonomy are a wide variety of complex performances that include such activities as projects, portfolios, demonstrations, experiments, fine art performances, and medical diagnosis or other professional practicum as well as teaching and extended group activities, discussions, and interviews. There are large bodies of work on such assessments, which some describe as *performance assessments*, although this term has multiple meanings and can refer to item types in more constrained categories, as well (Gronlund, 1982, 2003).

For assessment systems with considerable sophistication in the available scoring algorithms, it is sometimes possible to generate computer-based scoring for some aspects of assessment tasks in the presentation/portfolio category; however, there are many challenges to the validity of these scores and often human scoring or comparison human scoring is desirable.

For the purposes of this paper – innovations in computer-adaptive environments – only two things will be mentioned with regard to presentation-based assessments. First, computers can indeed facilitate preparation, collection, presentation and delivery of the products of such assessments, thereby enriching possible outcomes. For instance technology can make multimedia portfolios and interactive presentations operational.

Secondly, computers can create opportunities for innovation that involve group activity at distributed locations or at asynchronous times. Peer assessment is an example and so are activities that call for engaging in a community of practice. Knowledge is often embedded in particular social and cultural contexts. Although there is much need for assessment that involves students engaged with other students, there is limited knowledge of best practices for group assessment (Pellegrino, Chudowsky, & Glaser, 2001). Technology may offer some new tools.

A cautionary note about general computer-based presentation tasks that do not involve a component of human evaluation but instead rely only on machine algorithms for scoring: Building a case for why the automated scoring approach effectively measures the construct and properly scores the task(s) is extremely important. If these types of computer-based approaches are to be useful, developers must bring forward a great deal of external validity and reliability evidence for the scoring criterion comparisons. The evidence must be made available to system users, especially instructors and students who will be impacted by the scoring decisions and who need to understand the evaluations upon which the systems are functioning. Currently, the approach of many systems is to "black-box" the score approaches, such that it is difficult even for those with considerable measurement expertise to determine how the scoring systems are functioning and the degree to which the evidence is valid and reliable. Meaningful scoring is an important concern not only in large-scale testing but also in classroom products for assessment. It should be a standard for the field of education that e-learning developers, whether in large scale or classroom-based assessment, should incorporate and provide this evidence with the release of their products (Scalise & Wilson, 2006).

Conclusion

Assessment task design is a rich and complex arena in the rapidly emerging field of computer-based assessment and involves many considerations, including interactivity, the flow between items, assessment assembly specifications, and considerations of item feedback and learning interventions intermingled in the item flow. In this paper, we introduce a taxonomy of 28 item types – the *Intermediate Constraint Taxonomy for E-Learning Assessment Questions and Tasks* – that varies the student response format along a constraint dimension from completely selected to completely constructed. The many intermediate types between these two extremes are termed "intermediate constraint" items. Categories of the Taxonomy range from highly constrained types to fully open, and offer e-learning developers a variety of iconic formats to consider in designing assessment tasks. Strengths and weaknesses of various types are discussed and situated relative to each other. Examples are provided from our work in e-learning and assessment, with references in the Taxonomy table to other additional examples.

With the intent of consolidating considerations of item constraint for use in e-learning assessment designs, references for the Taxonomy were drawn from a review of 44 papers and book chapters on item types and item designs, many of which are classic references regarding particular item types. The 28 example types are based on 7 categories of ordering, involving decreasing the constraint on the response from fully selected to fully constructed, with four iconic examples in each category of constraint.

Many other innovative item formats can be derived from combinations or adjustments of the example types, and item formats across types can be widely varied depending on the domain to be measured and the inclusion of new media such as interactive graphics, audio, video, animation and simulation.

Mislevy (1996) makes the point that if the information provided by an innovative item type is no better than provided by conventional multiple-choice, then the innovation seems pointless. In other words, innovations must be justified by providing something beyond what is available through standard formats. For innovative assessment questions and tasks, what this "something" is might take many forms, from increasing predictive validity, to improving classroom effects, or to providing better metacognitive interventions by increasing the ability to diagnose paths to competency rather than simply ranking students.

Furthermore, Bennett (1993) notes that a high degree of constraint in the response does not necessarily preclude construction, which may be required by many multiple-choice tasks. But a criticism of multiple-choice has been that they are all too readily written to measure low-order skills that do not require significant construction.

Finally, as Osterlind (1998) maintains, "it should be emphasized that the new item formats, although more attractive in some respects than the common multiple choice format, are still required to meet psychometric standards of excellence" (p. 206). As the research base increases, one possible route to effective innovation in task design would be the complementary use of novel and conventional approaches to balance the possible advantages gained from new item types with the risk that comes from using innovations that are not yet well-researched.

Endnotes

1. College of Education, Educational Leadership (Applied Measurement), University of Oregon, Eugene, CA 97403
2. Cognition and Development: Policy, Organization, Measurement, and Evaluation; Education in Mathematics, Science & Technology (EMST), University of California, Berkeley, CA 94720
3. The front facet was used to develop the *Intermediate Constraint Taxonomy for E-Learning Assessment Questions and Tasks* in this paper (Table 1, page 9).
4. Osterlind (1998) states, "the term 'item' does not adequately cover the myriad formats that stimuli in the alternative approaches can assume. 'Exercises' may be a more generally appropriate descriptor; but, since almost everything described about test items also applies to the exercises, for practical purposes, the terms can be used interchangeably" (p. 204).
5. See Table 1, column 1 of the *Intermediate Constraint Taxonomy*.
6. *Ibid*.
7. *Ibid*.
8. *Ibid*.
9. See Table 1, column 2 of the *Intermediate Constraint Taxonomy*.
10. *Ibid*.
11. *Ibid*.
12. See Table 1, column 3 of the *Intermediate Constraint Taxonomy*.
13. *Ibid*.
14. *Ibid*.
15. *Ibid*.
16. The drop down choices for sentence one are "horizontally," "vertically" and "diagonally"; for sentence two they are "x-axis," "y-axis," and "both x-axis and y-axis"; and for sentence three they are "to the left," "to the right," "up" and "down."
17. See Table 1, column 4 of the *Intermediate Constraint Taxonomy*.
18. *Ibid*.
19. *Ibid*.
20. *Ibid*.
21. See Table 1, column 5 of the *Intermediate Constraint Taxonomy*.
22. *Ibid*.
23. *Ibid*.
24. *Ibid*.
25. See Table 1, column 6 of the *Intermediate Constraint Taxonomy*.
26. To make this example more compelling as truly incorporating the entire outcome space that might be constructed by a student, pairs of line segments as well as the entire line itself might also be offered as choices for selection.
27. See Table 1, column 6 of the *Intermediate Constraint Taxonomy*.
28. *Ibid*.
29. *Ibid*.
30. The response box should be enlarged for an extended essay. Students sometimes feel cued to the desired length for the essay by the size of the available entry box, even when extension sliders are available.

References

Attali, Y. & Burstein, J. (2004). *Automated essay scoring with e-rater V.2.0.* Paper presented at the Annual Meeting of the International Association for Educational Assessment, Philadelphia, PA.

Bennett, R. E. (1993). On the Meaning of Constructed Response. In R. E. Bennett, Ward, W. C. (Ed.), *Construction versus Choice in Cognitive Measurement: Issues in Constructed Response, Performance Testing, and Portfolio Assessment* (pp. 1–27). Hillsdale, NJ: Lawrence Erlbaum Associates.

Bennett, R. E., Goodman, M., Hessinger, J., Kahn, H., Ligget, J., Marshall, G., et al. (1999). Using multimedia in large-scale computer-based testing programs. *Computers in Human Behavior*, *15(3–4)*, 283–294.

Bennett, R. E., Sebrechts, M. M., & Yamamoto, K. (1991). *Fitting new measurement models to GRE general test constructed-response item data* (No. 89–11P GRE Board report, ETS research report 91–60). Princeton, NJ: Educational Testing Service.

Breland, H., Kukich, K., & Hemat, L. (2001). *GRE Automated-Editing Task Phase II Report: Item Analyses, Revisions, Validity Study, and Taxonomy Development* (Report No. GRE Board Professional Report No. 98–05P). Princeton, NJ: Educational Testing Service.

Chung, G. & Baker, E. (1997). *Year 1 Technology Studies: Implications for Technology in Assessment* (CSE Technical Report No. 459). Los Angeles: Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, University of California, Los Angeles.

Embretson, S. (2002). Generating Abstract Reasoning Items with Cognitive Theory. In S. Irvine, Kyllonen, P. (Ed.), *Item Generation for Test Development* (pp. 219–250). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Glaser, R. (1988). *Cognitive and environmental perspectives on assessing achievement*. Paper presented at the Assessment in the Service of Learning ETS Invitational Conference, Princeton, NJ.

Glaser, R. (1991). Expertise and Assessment. In M. C. Wittrock & E. L. Baker (Eds.), *Testing and Cognition*. Englewood Cliffs, NJ: Prentice-Hall.

Gronlund, N. E. (1982). Constructing Performance Tests. In *Constructing Achievement Tests* (pp. 81–92). Englewood Cliffs, NJ: Prentice-Hall, Inc.

Gronlund, N. E. (2003). *Assessment of Student Achievement, Seventh Edition*. New York: Pearson Education, Inc.

Haladyna, T. M. (1994a). Item Shells and Conclusions. In *Developing and Validating Multiple-Choice Test Items* (pp. 109–161). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Haladyna, T. M. (1994b). Measuring Higher Level Thinking. In *Developing and Validating Multiple-Choice Test Items* (pp. 87–107). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Haladyna, T. M. (1994c). Multiple-Choice Formats. In *Developing and Validating Multiple-Choice Test Items* (pp. 35–57). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Leacock, C. (2004). Scoring Free-Responses Automatically: A Case Study of a Large-Scale Assessment. *Examens*, *1(3)*.

McDonald, M. E. (2002). Developing Multiple-Choice Items. In *Systematic Assessment of Learning Outcomes* (pp. 83–120). Sudbury, MA: Jones and Bartlett Publishers.

Mislevy, R. J. (1996). Test Theory Reconceived. *Journal of Educational Measurement, 33*, 379–417.

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

Page, E. B. & Petersen, N. S. (1995). The computer moves into essay grading. *Phi Delta Kappan*, *76*, 561–565.

Parshall, C. G. (2002). Item Development and Pretesting in a CBT Environment. In C. Mills, Potenza, M., Fremer, J., Ward, W. (Ed.), *Computer-Based Testing* (pp. 119–141). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Parshall, C. G., Davey, T., & Pashley, P. J. (2000). Innovative Item Types for Computerized Testing. In W. Van der Linden, Glas, C. A. W. (Ed.), *Computerized Adaptive Testing: Theory and Practice* (pp. 129–148). Norwell, MA: Kluwer Academic Publisher.

Parshall, C. G., Spray, J., Kalohn, J., & Davey, T. (2002). Issues in Innovative Item Types. In *Practical Considerations in Computer-Based Testing* (pp. 70–91). New York: Springer.

Pellegrino, J., Chudowsky, N., & Glaser, R. (2001). Knowing What Students Know: The Science and Design of Educational Assessment. In N. R. C. Center for Education (Ed.). Washington, D.C.: National Academy Press.

Resnick, L. B. & Resnick, D. P. (1992). Assessing the Thinking Curriculum: New Tools for Educational Reform. In B. R. Gifford & M. C. O'Connor (Eds.), *Changing Assessments: Alternative Views of Aptitude, Achievement and Instruction* (pp. 37–76). Boston, MA: Kluwer Academic Publishers.

Rijmen, F. & De Boeck, P. (2002). The random weights linear logistic test model. *Applied Psychological Measurement*, *26*, 217–285.

Scalise, K. (2004). BEAR CAT: *Toward a Theoretical Basis for Dynamically Driven Content in Computer-Mediated Environments*. Dissertation University of California, Berkeley.

Scalise, K. & Wilson, M. (2006). Analysis and Comparison of Automated Scoring Approaches: Addressing Evidence-Based Assessment Principles. In D. M. Williamson, I. J. Bejar & R. J. Mislevy (Eds.), *Automated Scoring of Complex Tasks in Computer Based Testing*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Shavelson, R. J. (2001). On Formative (Embedded) Science Assessment. On *Great Minds Program Dialogue, Science Education in the 21st Century: Pushing the Envelope on Student Assessment*. Stanford University.

Shavelson, R. J., Black, P., Wiliam, D., & Coffey, J. (2003). On *Aligning Formative and Summative Functions in the Design of Large-Scale Assessment Systems*. (in progress).

Shepard, L. (1991a). Interview on assessment issues with Lorrie Shepard. *Educational Researcher*, *20(2)*, 21–23, 27.

Shepard, L. (1991b). Psychometricians' beliefs about learning. *Educational Researcher*, *20(6)*, 2–16.

Stout, W. (2002). Test Models for Traditional and Complex CBTs. In C. Mills, Potenza, M., Fremer, J., Ward, W. (Ed.), *Computer-Based Testing* (pp. 103–118). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Van der Linden, W. (2002). On Complexity in CBT. In C. Mills, Potenza, M., Fremer, J., Ward, W. (Ed.), *Computer-Based Testing* (pp. 89–102). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Author Biographies

Kathleen M. Scalise, Assistant Professor, University of Oregon, Applied Measurement and Instructional Design, Educational Leadership. Dr. Scalise is interested in analysis of student learning trajectories with item response models, computer-adaptive approaches for assessment, dynamic delivery of content, and innovative instructional design practices in science and mathematics education. In recent projects, she has served as science writer for Curriculum Frameworks and Instructional Resources Division of the California Department of Education for the 2002 California Science Framework for K–12 Public Schools, and on the ChemQuery and Living by Chemistry (NSF) projects at the University of California, Berkeley.

Email: kscalise@uoregon.edu

Bernard Gifford, Professor, University of California, Berkeley, Engineering, Mathematics, Science and Technology Education. Dr. Gifford is a specialist in mathematics education and president/ chief instructional designer of the Distributed Learning Workshop, which develops e-learning instructional materials in mathematics for a variety of venues, including the University of California Advanced Placement Online Courses in Calculus. Gifford has a long history of scholarship and service in the field of education. He was deputy chancellor of the New York City Public Schools from 1973–77, and he served as dean of Berkeley's Graduate School of Education from 1983–89. He was also vice president for education at Apple Computer from 1989–92. His numerous books include *Policy Perspectives on Educational Testing* (1993) and *Employment Testing: Linking Policy and Practice* (1993).

Email: bgifford@dlworkshop.net

J·T·L·A

The Journal of Technology, Learning, and Assessment

www.jtla.org

Technology and Assessment Study Collaborative
Caroline A. & Peter S. Lynch School of Education, Boston College