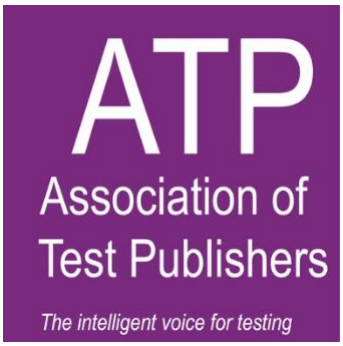


Innovative Item Types

A White Paper & Portfolio

Association of Test Publishers (ATP) and the

Institute for Credentialing Excellence (ICE)

Alternative Item Types

DISCLAIMER: This White Paper is intended only to provide general, high-level guidance concerning the use of innovative type items in the credentialing space. Each reader must consider whether any given section or subsection is applicable to his or her specific program(s), credential(s), or test(s).

While the ICE and the ATP have made every effort to ensure that the information contained in this document has been developed from reliable sources, all information is provided “as is” and neither the ICE, the ATP, nor any of the participating publishers or service providers, makes any warranty, express or implied, nor do they collectively or separately assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, product, or process described in this document. In no event will the ICE, the ATP, their agents, or employees, be liable to any user of this document for any decision made or action taken in reliance on the information in this document, including but not limited to liability for any consequential, special or similar damages, even if the ICE and the ATP were advised of the possibility of such damages.

The information in this White Paper is provided with the understanding that neither the ICE nor the ATP, nor any individuals who participated in the preparation of this document, shall be deemed to be engaged in rendering legal, technical, psychometric, or assessment advice and services. Therefore, this document should not be used as a substitute for consulting with competent legal, technical, and measurement specialists.

Table of Contents

Introduction

This project, which includes both a white paper and an accompanying portfolio of sample items, is the result of collaborative efforts by the Institute for Credentialing Excellence (ICE) and the Association of Test Publishers (ATP). The paper provides an overview of considerations and best practices for incorporating alternative item types into an assessment. The focus of the paper is on credentialing/certification assessments, but many of the same considerations and processes apply to other types of examinations (i.e., assessments for use in educational, industrial/organizational, or clinical settings). For simplicity, this paper will use the term "credentialing organizations" to refer to both certification and licensing organizations.

The paper first defines alternative item types and discusses potential benefits of using them within an assessment program. The bulk of the paper details best practices for evaluating the appropriateness and feasibility of incorporating alternative item types into your assessment, including validity evidence requirements, design and development best practices, test security or memorability concerns, Americans with Disabilities Act (ADA) compliance issues, administration, scoring considerations, examinee preparation, and cost and resource considerations.

The portfolio of sample alternative items is provided to inform test developers, examination committees, or other stakeholders, by illustrating how various constructs can be measured using several commonly used, as well as customized, item formats. The sample items in the portfolio have been donated by sponsors of certification examinations and education assessments and testing vendors. Please note that ICE and ATP do not endorse any particular item type; instead we encourage credentialing organizations or other types of test sponsors to conduct appropriate research necessary to ensure any alternative item types under consideration are appropriate and will positively contribute to the assessment's measurement properties.

What are innovative or alternative items?

In this paper, the terms "innovative" or "alternate item types" (AITs) are used to describe any items that differ from the traditional multiple choice item type (MCIT). The MCIT is an item which contains a text-based question and answer options (most commonly four options) and a single correct answer. By comparison, AITs can be in any format—computer-based (e.g., hot spot or drag-and-drop), paper-and-pencil (e.g., essay or short answer, situational judgment tests), or performance tasks (e.g., verbal language or translation tasks, ear mold, surgical skills test). The primary focus of this paper is on computer-based or technology-enhanced items. Not only do computer-based examinations represent the largest sector of the testing industry, but many item type innovations have stemmed from advances in technology, especially computer technology – and in our view, many other sponsoring organizations may be considering using computer-based items in the future.

Several different systems or taxonomies for classification of AITs have been proposed in the research literature (Scalise & Gifford, 2006; Sireci & Zenisky, 2006; Parshall, Spray, Kalohn, & Davey, 2002; Parshall, Harmes, Davey, and Pashley, 2010). These taxonomies are useful in understanding the ways in which items differ from one another. For example, one simple taxonomy classifies items into two categories—selected response and constructed response items. Selected response items require the examinee to select one or several correct responses

from a pre-determined list of response options (e.g., multiple choice). Conversely, constructed response items require the examinee to supply a response to the item (e.g., fill-in-the-blank).

Many of the taxonomies are not sufficiently complex or robust to capture the range of possible differences between AITs. While no taxonomy works in all situations, the seven-dimension taxonomy by Parshall, Harmes, Davey, and Pashley (2010) seems particularly useful in highlighting the various ways items can differ from one another, so it has been chosen for this discussion. That taxonomy is provided below, along with some questions that test developers should consider under each dimension. Parshall et al. (2010) note that each dimension represents a continuum and there are decisions that test developers must make in relation to each dimension when designing and developing AITs and their associated interfaces. Please note that many items are innovative on more than one dimension because the dimensions may not be independent from one another (e.g., adding media to an item may also change how an examinee responds to the item).

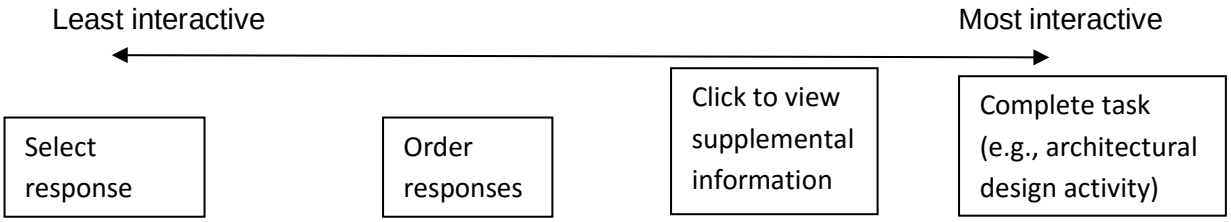
1. ***Assessment/item structure*** is the structure of item presentation and type of response required of the examinee. What is the mode of presentation of the item(s) to the examinee (e.g., computerized or paper)? What is the nature of the response elicited from the examinee? Is the examinee being asked to select from a set of provided responses (e.g., multiple choice) or to create, construct, or synthesize a response (e.g., essay)? The item types should be selected to improve measurement of critical constructs related to the assessment.

2. ***Complexity*** is the number and variety of elements that an examinee must consider when responding to an item. Is the number of tasks appropriate given the intended construct to be measured? Does the complexity of the item have deleterious effects on the ability to score the item? Do examinees have the necessary computing skills to process, interact with, and respond to the item, and are tutorials provided? Effort should be made to ensure more complex items (and the interface required for their delivery) are consistent with the inferences the credentialing organization wants to make and do not introduce construct irrelevant variance or error into the assessment. AITs also have the potential to add psychometric complexity for the testing organization (e.g., scoring, comparability, equating, standard setting). The scoring of AITs is discussed in more detail later in this paper.

3. ***Fidelity*** is the degree to which the assessment provides a realistic and accurate reproduction of actual objects, situations, tasks, or environments that are familiar to the examinee's daily experience and are a part of the construct being measured. To what extent does the assessment task reflect the skill, task, or behavior that we are most interested in observing in the examinee? What degree of fidelity to the "real-world" environment is necessary to elicit these examinee behaviors? Is software adequate, or is a hands-on, physical representation necessary? *Higher fidelity does not mean the assessment will be more valid.* The fidelity level should also closely resemble the score inferences that the credentialing organization wants to make. Sponsors are cautioned that although higher fidelity items may be valuable within the assessment process, higher fidelity items may not always be fair to all examinees (e.g., all examinees may not be familiar with environment in the item), may be unnecessarily complex, or may create challenges for examination administration.

4. ***Interactivity*** is the extent to which the item responds or reacts to examinee inputs. Does the item reflect single or multiple stages? Are the steps discrete or continuous?

Does the examination provide some degree of feedback to the examinee (e.g., simulated patient response to inquiry, results of selected laboratory tests or diagnostics, changes in graphical presentation, additional instructions, or branching)?



Extensive design and development are required for the highly interactive items, which includes determining how to score them. Organizations using highly interactive items (e.g. simulations) may need to include constraints in the item to prevent examinees from continuing too far down a path of incorrect responses.

5. ***Media inclusion*** is the incorporation of graphics, photographs, audio, animations, or video to an item to expand measurement of a construct, more faithfully reflect real-world environments, reduce unnecessary dependence on reading skills, and potentially increase validity of scores. The inclusion of graphics is common in AITs (e.g., hot spot, plotting, medical images). Audio is mostly used in assessments measuring language skills or music; however, there may be opportunities in other areas in which the processing of aural information is critical to a task. Videos seem to be beneficial to measure interpersonal communication, aspects of human interactions, or dynamic processes or movement. The possible challenges with media inclusion are file sizes, file types, fairness to examinees who need special accommodations with visual or audio components, memorability of items (i.e., test security), high cost of editing and production, and the potential for adding construct irrelevant variance (especially in videos).

6. ***Response action*** is the physical action required of the examinee to respond to an item and the input device used. What is the examinee being asked to do in the question (e.g., typing on a keyboard, clicking a mouse to select answer, clicking a mouse to select object then dragging it to another location on screen, touching a screen to select a response)? It is critical that the required response actions be consistent with capabilities of the target audience for assessment and relevant to the construct being measured, and that examinees are given clear instructions and/or tutorials on how to respond to items.

7. ***Scoring methods*** are processes for converting examinee responses into a quantitative score. What are the scoring opportunities or events in the item? If the tasks or behaviors are multi-faceted or otherwise complex, how will they be divided and reduced to discrete scorable pieces? Is the response simple enough to be scored automatically, or is some element of human judgment required? Automated scoring (versus manual scoring) is frequently required or desirable for computer-based assessments. There are many options for scoring, from dichotomous scoring (correct/incorrect) to partial credit or weighted scoring models to complex modeling (e.g., those associated with emulations and simulations). The scoring for an AIT should be developed in conjunction with the assessment with input from a psychometrician or measurement professional, and

designed to be consistent with the inferences that the credentialing organization wants to make from the test scores.

It is important to note that the sample items in the accompanying portfolio of AITs are organized by item format (ordered by relative popularity/availability) since credentialing organizations typically opt to use the item formats offered by many testing vendors (e.g., drag-and-drop, hot spot) rather than incur the expense of creating a customized item format.

Why use alternative items?

Although there is relatively little research documenting the benefits of AITs, the appeal of AITs stems from beliefs that these items: (1) measure the intended constructs better than traditional MCITs; (2) measure constructs that could not be assessed by MCITs; (3) increase the measurement precision (i.e., reliability) of the assessment; (4) increase the measurement efficiency of the assessment; (5) increase the fidelity or face validity of the assessment with respect to the actual functions that the examinee performs in daily role/job/practice (without sacrificing construct validity or reliability); and/or (6) measure higher order thinking or cognitive functioning better than MCITs. The items in the accompanying portfolio provide examples of how some test sponsors are expanding or improving the measurement of constructs with commonly used and customized AITs. A few examples from the portfolio are provided later in this paper.

As mentioned above, the body of empirical research focusing on the relative effectiveness of AITs is growing, but still relatively small. A summary of the available research is provided below.

- Research studies on validity and/or reliability and other psychometric properties, including the following:
    - Downing, Baranowski, Grosso, and Norcini (1995) compared traditional MCITs and multiple true-false (MTF) items in a medical certification exam. They found MTF items were more reliable than MCITs. However, in a criterion-related validity study, the scores from MCITs had higher correlation with independent performance ratings than the MTF items. Both item types were found to measure similar cognitive abilities.
    - Collins and Waugh (2008) found that multiple response and brief constructed response items performed similarly to traditional MCITs in terms of difficulty level; although the AITs had slightly better item-level reliability.
    - Wan and Henly (2012) investigated the reliability and construct validity of two groupings of AITs on a K-12 science test. The authors investigated figural response items (which were items that included an illustration, graph, or diagram and required examinees to select regions of the figure, complete a figure by dragging and dropping elements, or reordering elements of the figure) and constructed response items (which were items that required examinees to type in a response of between one sentence and a few sentences in length). Using confirmatory factor analysis, the researchers found that the figural response and constructed response items measured similar constructs to traditional MCITs. Using item response theory (IRT) information functions, the researcher found the figural response items

provided a similar amount of information as traditional MCITs, but the constructed response items provided more information than traditional MCITs.

- Woo, Kim, and Qian (2014) investigated the psychometric properties of three types of AITs (fill-in-the-blank calculation, multiple response, and ordered response) compared to traditional MCITs. These items were administered as part of a computer-adaptive nursing examination. Woo et al. found that the multiple response items were the most difficult of the four item types and the fill-in-the-blank calculation items were the easiest. There was a significant difference between each AIT and the traditional MCIT. Fill-in-the-blank calculation items were most discriminating, but both fill-in-the-blank calculation and multiple response items were significantly more discriminating than traditional MCITs. Fill-in-the-blank calculation items were the most difficult to guess the correct answer, but both fill-in-the-blank calculation and multiple response items were significantly more difficult to guess the correct answer than traditional MCITs. Fill-in-the-blank calculation items also provided more information than the other AITs and traditional MCITs. By analyzing the cognitive classifications of the items (e.g., Knowledge, Comprehension, Application, and Analysis), traditional MCITs were found to assess higher order thinking skills than fill-in-the-blank calculation, multiple response, or ordered response items. When investigating item drift, they found that traditional MCITs become easier over time and multiple response items become more difficult over time. Fill-in-the-blank calculation and ordered response had no significant drift over time.

  Woo et al. (2014) also compared simple text MCITs with items that included graphics, audio, exhibits, or graphics and exhibits. They found items that included graphics were significantly easier than simple text MCITs. There were no significant differences in discrimination between simple text-based items and items with graphics, audio, exhibits, or graphics and exhibits. Items with exhibits or exhibits/graphics were best at assessing higher order thinking skills, and these items as well as items with audio assessed higher order thinking skills better than simple text MCITs. Conversely, items with graphics assess lower order thinking skills compared with simple text MCITs. In terms of drift, items with graphics became easier over time. There was no significant drift for items with audio, exhibits, exhibits/graphic, or simple text MC items.

- Krogh and Muckle (2017) found there was not a significant difference in performance on AITs compared to MCITs for most examinees; only a small minority of examinees (6.7%) exhibited a significant difference in performance between these item types. Examinee scores (in the form of Rasch ability estimates) on MCITs and AITs exhibited a fairly high correlation of r=0.58. While significant differences in item difficulty were observed among some of the individual item formats, the aggregate of AITs exhibited a comparable item difficulty to the MCITs; therefore, the average difficulty level of the examination remained similar. The AITs items took significantly more time to answer than MCITs but were more discriminating. The AITs exhibited comparable dimensionality to MCITs, and a unidimensional IRT model was deemed appropriate for analyzing both AITs and MCITs. The new item types

were found to have acceptable attributes for inclusion in the certification program's high-stakes examinations.

- Research studies on assessment efficiency (i.e., how much time it takes for an examinee to respond to an item), including the following:

    - Jodoin (2003) compared IRT information for MCITs and two AITs (drop-and-connect and create-a-tree) used on the Microsoft Certified System Engineer examination. Both AITs provided more information across all ability levels than MCITs, but it took longer to respond to the AITs compared to MCITs.

    - Wan and Henly (2012) investigated efficiency of the figural response (defined in above section) and constructed response (defined in above section) items compared to traditional MCITs. They found that examinees required a similar amount of time to respond to the figural response items and MCITs. Examinees required more time to respond to the constructed response items than either of the other two.

    - Woo, Kim, and Qian (2014), as part of the study described in the previous section, also compared the length of time spent on fill-in-the-blank calculation items, multiple response items, and ordered response items to the time spent on MCITs. Fill-in-the-blank calculation items took the most time for examinees to respond, but all three AITs took significantly longer than MCITs. When comparing simple text MCITs to items with graphics, audio, exhibits, or graphics and exhibits, they found that items with exhibits took the most time, but items with exhibits, audio, and graphics/exhibits also took significantly longer than simple text MCITs. Conversely, items with graphics took significantly less time than simple text MCITs.

    - Dwyer, Penny, and Johnson (2015) compared the average testing time of traditional MCITs, multiple-choice multiple response items, and drag-and-drop items. They found that, on average, it took examinees 58% longer to respond to multiple-choice multiple response items than traditional MCITs and it took examinees nearly 200% longer to respond to drag-and-drop items compared to traditional MCITs.

Considerations in Selecting, Developing, and Implementing Alternative Items

For new credentialing programs or existing programs considering a transition to AITs, there are many questions to consider in order to evaluate whether AITs will benefit the assessment program. Among those considerations are:

- What is the purpose of your program? What is the mission of the credentialing organization: to protect the public or to verify that educational standards have been met? The purpose of the organization sponsoring the assessment should be the starting point for assessment design. Those charged with governance of high-functioning

organizations have a deeply-ingrained understanding of why they exist in the first place, and this philosophy infiltrates all of the activities undertaken by the organization.

- What are you trying to measure and what are the appropriate ways to measure it? What knowledge or skills are we trying to authenticate or identify in our population of interest? What are the most effective means of measuring these with precision? What decisions will be made with the data collected from these types of items? There should be awareness of how knowledge or skills are evaluated in practice and how they can be captured in (or adapted to) a standardized assessment, as well as how the assessment data will be interpreted and used by organizational stakeholders.

- Do you have sufficient resources (e.g., financial, technological, human) for the development, administration, and scoring of AITs? Which innovations are realistic, and which lack feasibility? Visionary ambitions must sometimes be tempered by realistic limitations. Organizations must take into account their own strength, influence, and assets, as well as how they may be harnessed and leveraged to achieve measurement goals. Organizational governance involves a duty of care and a fiduciary obligation to channel resources effectively and responsibly.

Is there a process for designing and developing alternative items?

Parshall and Harmes (2008) proposed a six-step process for the design of AITs. This process provides a helpful structure for an organization that is considering incorporating AITs into its assessment. The process also helps to set realistic expectations—developing and implementing AITs is not a quick and easy process and requires iteratively refining and evaluating the items. Parshall and Harmes' process steps are listed below. Included with the descriptions are resources for additional information on the topics.

1. **Analyze the exam program's construct needs** to determine strengths and weaknesses or omissions in the current assessment. This analysis involves evaluating how well the assessment aligns to its purpose and measures the knowledge and skills identified by the test blueprint (e.g., whether the assessment measuring the declarative or procedural knowledge when the skill to do a complex task supported by that knowledge would be a better measure of examinee ability). The Wendt and Harmes (2009a) article documents the National Council Licensure Examination (NCLEX) process for identifying areas in which AITs could enhance the measurement of the constructs associated with its exam program. The NCLEX approach may provide insights to an organization at this step in the process.

2. **Select specific innovations for consideration** that may enhance measurement for weak or missing areas identified in Step 1. In "Designing Templates Based on a Taxonomy of Innovative items," Parshall and Harmes (2007) provide a table (Table 1) that lists various types of innovation (based on the taxonomy of dimensions noted in the introduction of this paper) and associated advantages and challenges of each dimension, which may provide useful guidance in this step of the process.

3. **Design initial prototypes** by having test developers and subject matter experts (SMEs) define item types based on selected innovations and draft an initial design, including potential scoring protocols. These initial draft prototypes should be reviewed by internal exam program stakeholders and refined as needed before moving onto step 4. At this

step, it may be helpful to review some sample AIT templates provided in "Improving the Quality of Innovative Item Types: Four tasks for design and development" by Parshall and Harmes (2009).

4. **Iteratively refine item type designs** through the tasks listed below. This set of activities is the most extensive in this model:
    a. Develop item writing materials and sample items;
    b. Conduct usability testing on the sample items; and
    c. Evaluate and revise item type designs.

5. **Pilot test alternative item types**, which should include all phases in the item and examination life cycle (i.e., item banking, test publishing, test delivery or administration, examinee response capturing, item analysis, and scoring).

6. **Produce final materials** that will be needed to implement the new item types. This includes exam information for examinees (e.g., candidate handbook, tutorials, website), item writer training information, scoring rubrics, and rater training materials if manual scoring is needed.

Steps 1 and 2 in the above process are designed to ensure that critical thought is given to the potential effects of AITs on the validity and reliability of the exam. As mentioned in the introduction, when deciding what item formats to use, an organization must consider how well different item types measure the intended construct(s) (e.g., job-related competency). As Jones and Vickers (2011) stated, "The validity of inferences being made about scores must be based on valid, reliable, and fair assessments" (p. 4). In addition to validity considerations, the way AITs are weighted and scored will likely impact the reliability of the exam (i.e., the precision/reproducibility of test scores). Scoring-related issues will be addressed in more detail later in this paper.

Optimizing construct representation

There are several ways in which AITs may improve validity, such as increasing predictive validity, better representing job/role/practice content, and reducing construct irrelevant error. For example, modifying the presentation of items through more visual displays, such as graphics, will reduce reliance on reading ability and cognitive load, which may not be relevant to the performance domain to be measured (Strain-Seymour, Way, & Dolan, 2009). However, if the innovation is computer-driven, the examinee population's computer skills should also be taken into account (Parshall & Harmes, 2007; Sireci & Zenisky, 2006). For professions that require little in the way of computer skills, complex, computer-based innovations in testing may introduce construct-irrelevant variance.

AITs may also measure a broader array of skills and ability more easily than MCITs. For example, MCITs measure declarative knowledge (e.g., knowledge of the pieces, processes, and preferred approaches) rather than proficiency in performing tasks (Huff & Sireci, 2001, cited in Strain-Seymour et al, 2009). If a program wants its examination to measure higher order skills (e.g., analysis, skill, or motivation), AITs may be beneficial (Knapp, 2004; Jones & Vickers, 2011). Health, legal, and intelligence professionals must possess a large body of technical knowledge that can be well measured with MCITs. They also require the ability to analyze large amounts of information, identify critical issues, and recommend courses of action, where AITs may be able to assess the abilities needed to successfully do this. Physician medical licensing

tests, for example, use assessment methods based more in performance, in which examinees must recommend tests, interpret results, and suggest diagnoses based on patient examinations and lab results (Parshall & Harmes, 2007).

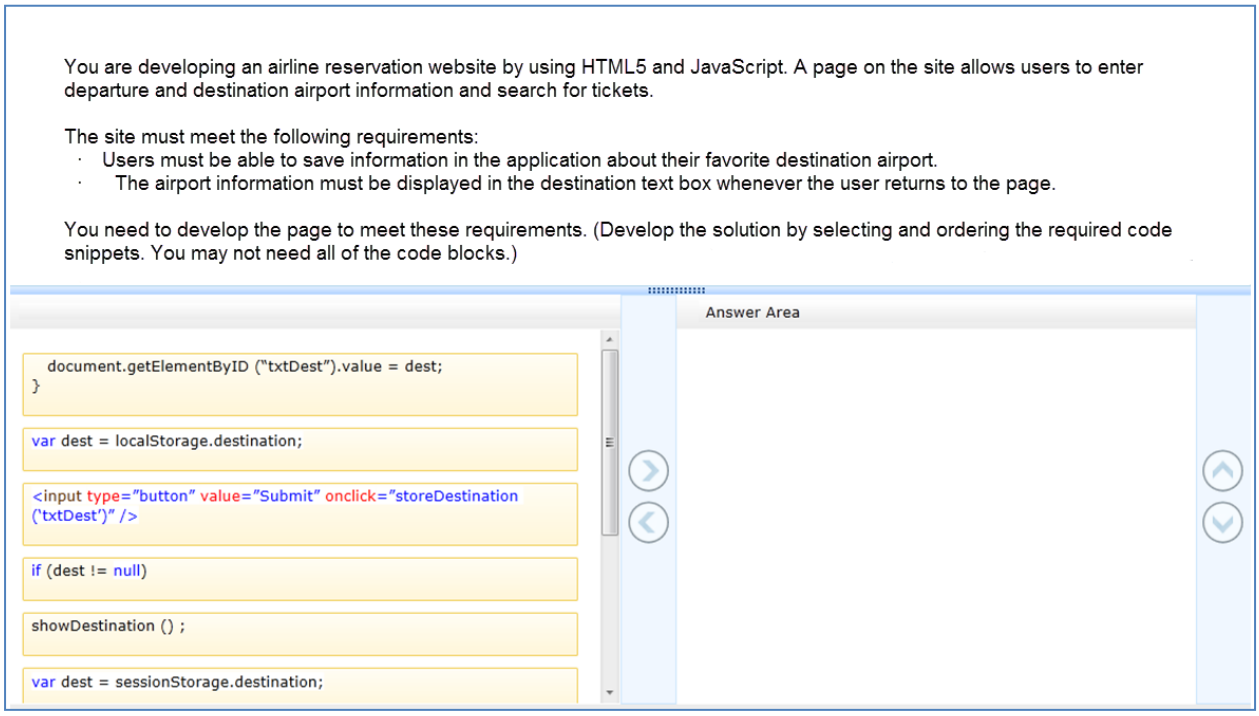
Sample AITs

To better evaluate the use of AITs, five samples are presented below with a description of the item and potential measurement gains or other benefits of using them.

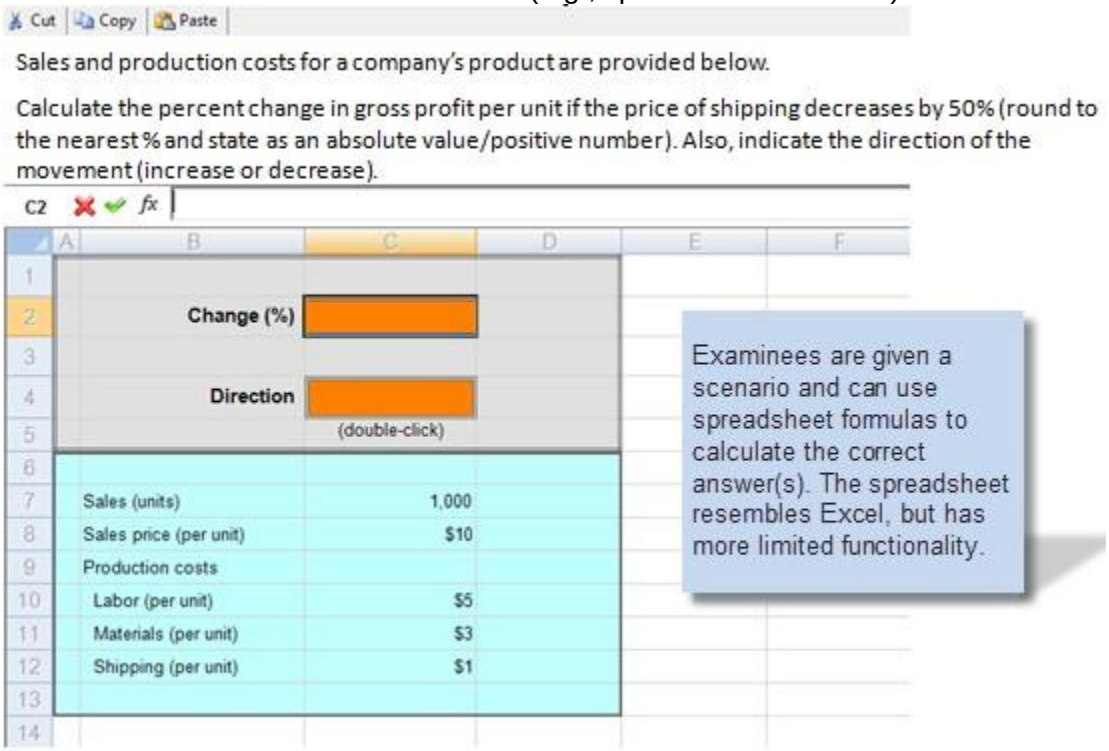
Sample 1: This item requires examinees to recognize skin conditions and identify which condition can appropriately be treated with cryotherapy. In this item, graphics likely reduce the cognitive load of an equivalent traditional MCIT that describes the condition in each photo. Additionally, this item increases the fidelity of the assessment.

Click on the condition that an adult-gerontology primary care nurse practitioner appropriately treats with cryotherapy.

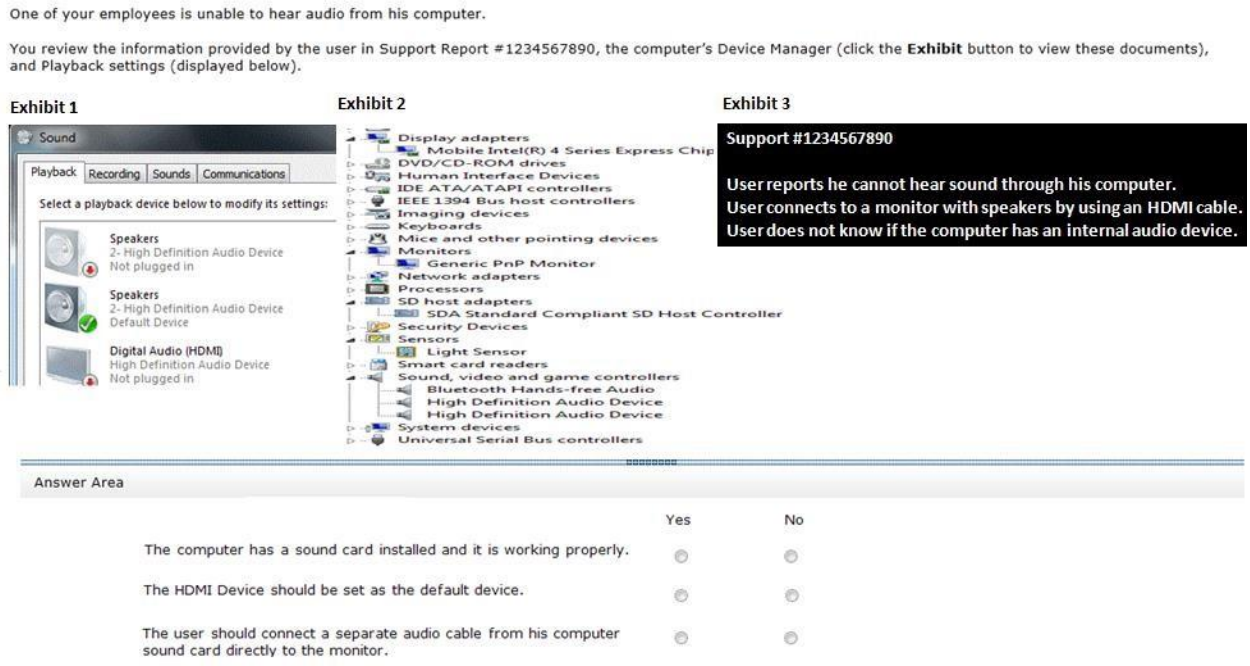
Sample 2: This item requires examinees to select the appropriate HTML and JavaScript code sections and put them in the correct order for developing a website that meets the provided scenario and requirements. This item provides more fidelity than a traditional MCIT and allows the credentialing organization to measure an examinee's ability to develop code and still score the item in an automated manner.



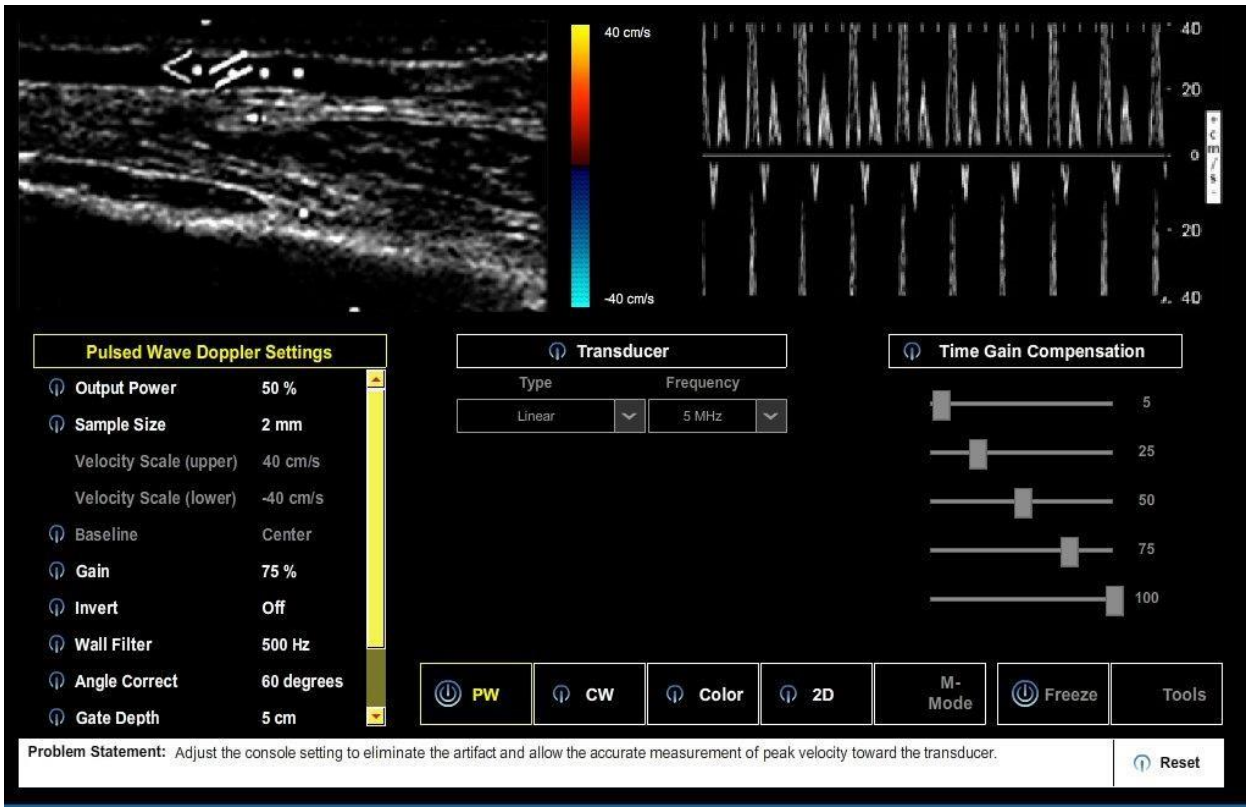
Sample 3: This item includes a worksheet that has similar functionality to Excel. The item requires the examinee to use the spreadsheet functionality to calculate answers or portions of answers to a provided financial scenario. Examinees can use any of the blank cells in the spreadsheet to calculate the answer. This item provides higher fidelity and measures more than could be measured in a traditional MCIT (e.g., spreadsheet functions).



Sample 4: This item requires the examinee to use the provided exhibits and information to determine whether three statements related to troubleshooting a computer audio issue are true or false. The item has more fidelity, is more interactive, and is more complex than a traditional MCIT. This item likely improves measurement of the job-related constructs.



Sample 5: This example is a custom, semi-interactive console item that measures an examinee's skill to solve an identified problem with ultrasound machine settings. This item contains two static images at the top of the screen. Below the images is an interactive mock-up of an ultrasound machine console. At the bottom of the screen is the problem statement, which asks the examinee to adjust the console settings to eliminate the artifact and allow accurate measurement of peak velocity toward the transducer. The examinees can click on and modify many of the console settings, however, images do not change when console settings are adjusted. For scoring purposes, allowances are made for personal preferences, but if they choose an incorrect setting or one that would downgrade the image, points are deducted. Partial credit is awarded.



As should be apparent from the above samples, AITs may better measure what examinees do not know, in addition to what they do know. For example, in a drag-and-drop scenario, where an examinee mistakenly places a stimulus may help organizations identify a performance gap that might direct future training courses.

Further, AITs may increase validity by reducing successful guessing. This can be managed by introducing complexity into selected response items (e.g., matching, rank order) and eliminating choices by replacing MCITs with brief constructed response items (Collins, Keenan, & Ramli, 2008; Sireci & Zenisky, 2006; Strain-Seymour et al. 2009).

Adhering to the test blueprint

Credentialing programs demonstrate content validity by identifying test blueprint specifications through a rigorous job/role/practice analysis to ensure the test content reflects content related to the job/role/practice domains. The job/role/practice analysis is the foundation for demonstrating the validity of the assessment; but the concept of validity is defined in the *Standards for Educational and Psychological Testing* (2014) as "the degree to which accumulated evidence and theory support a specific interpretation of test scores for a given use of a test" (p. 225). Therefore, a program that wishes to incorporate AITs into its assessment(s) needs to ensure those items contribute to validity evidence given the intended use of the test scores and constructs to be measured in the assessment(s). To support this decision making process, it is helpful to have discussions with SMEs and stakeholders during the job/role/practice analysis phase regarding the appropriate cognitive level of difficulty/complexity for measuring each

critical knowledge or skill. A commonly used taxonomy for classifying cognitive difficulty/complexity by test developers is a condensed version of Bloom's (1956) taxonomy, composed of three classifications—Recall/Recognition, Understand/Apply, Analyze/Evaluate. It may also be helpful to discuss what an examinee should demonstrate as evidence that the examinee possesses the desired level of the critical knowledge and skills. It is important to review the critical tasks from job/role/practice analysis and ensure the AITs are in line with these tasks and measure the breadth and depth of the job/role/practice covered by the credential.

Some AITs (e.g., scenario-based item sets, situational judgment test items) may assess multiple content domains and levels of the cognitive domains. Furthermore, how items are scored and weighted may vary from item to item or item type. Therefore, a program must be careful that examination forms constructed with a mix of AITs and scoring schemes will provide an equivalent experience to examinees, including equivalent construct representation and an equivalent passing score.

Implementing a mixed format examination may require that the testing program add rules or specifications to the test blueprint to address the resulting complexities of this situation. Some questions that may need to be answered include:

- Are the exam domain weights based on number of items or number of possible points?
- Must each exam form contain the same number of items of each format?
- Will the distribution of item formats be consistent across exams, such that each domain contains the same number of item types, items of each weighting, items of each scoring schema, etc.?
- If the content of an AIT spans exam domains and is worth more than one point, can the item be classified in two exam domains in the blueprint and subsequently be counted twice in the blueprint? Does the exam duration (i.e., time allotted for an examinee to complete the exam) need to be evaluated for each new exam form?

While it is not a requirement to keep item formats static across forms, changes in the distribution of item formats from examination form to examination form must not change the extent to which the test content maps onto the test blueprint specifications, create inequities in examination difficulty across forms, or change the examinee requirements (e.g., time needed to review and respond to items). Varying scoring schemes and mixed format assessments do not have to be complicated and can best capture the full domain of knowledge and skills a certification exam is intended to measure.

Programs that transition to different item formats should compare examinee performance on AITs (and delivery methods) to performance on the traditional MCITs and delivery methods (Krogh & Muckle, 2017). They should also examine reliability indices (e.g., whether the test consistently measures the construct). It is already well established that computer-based tests have demonstrated equivalent validity to paper-based versions while being more efficient to administer (McBride & Martin, 1983, cited in Sireci & Zenisky, 2006).

Increasing credibility with examinees and other stakeholders

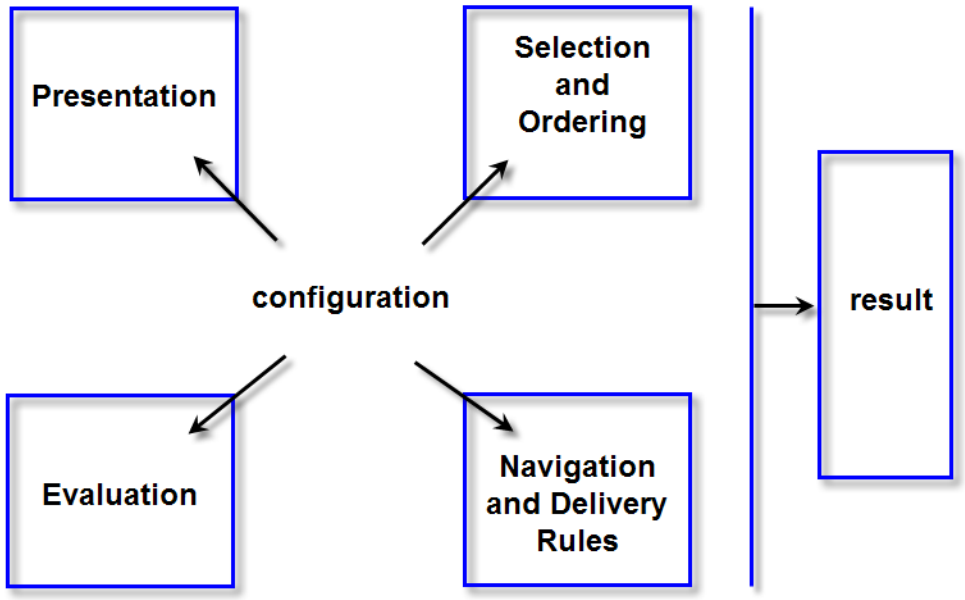
One advantage of AITs is their higher authenticity or fidelity, such that they more closely resemble "real life" contexts in which examinees would solve problems (Knapp, 2004; McSweeney, 2013). For example, a program intended to certify computer programmers might want to ask examinees to write, test, and/or correct a coding program as part of the assessment. A certification program for phlebotomists might want examinees to "drag-and-drop"

an animated needle onto an animated arm to identify the best location to make a blood withdrawal since it is not feasible to have examinees actually draw blood during an assessment. The face validity of these types of items may even increase the perceived credibility of the assessment and program. While face validity is no longer considered a psychometric concept, it is never-the-less an important marketing concept and a program's brand and credibility depends on, among other things, the extent to which examinees find a voluntary certification exam credible. If AITs can increase face validity without compromising validity in a psychometric sense, they are beneficial.

Technology considerations

There are additional technology considerations when designing items for computer-based delivery. Ramstad (2013) has suggested a model for AIT design that blends technology with sound design principles. The model is depicted in Figure 1 below, followed by questions to consider at the design phase for each component (Ramstad, 2013).

Figure 1. Components of AIT Design



Presentation: What is presented on the screen? How is the response made? Why is a particular modality more or less appropriate than another?
Selection and Ordering: Is the item part of a case study or larger set of items in a scenario? Does it have to precede or follow something else? Does it need to be grouped together with other content? On what basis is it selected for delivery to the examinee?
Navigation and Delivery Rules: What does the examinee have to do to the item to consider it "complete"? What other tools and resources must be provided to the examinee with the item?
Evaluation: What makes the item "correct"? How is a score derived? How are item scores used in total or section scoring and reporting?
Configuration: Once a new AIT has been designed, questions of configurability and usability should be considered. What aspects of the item can be changed to create variants?
Results: What must the results data include to be meaningful? How will the data be used?

The components and questions described in the model in Figure 1 are usable during steps 3 and 4 of the AIT design process proposed by Parshall and Harmes (2008), described earlier in this paper.

Resource Considerations

The good news is that the overall process for developing traditional MCITs can also be followed (including using the same SMEs) when developing AITs. The tools may be different, and in the case of technology-driven innovations, the coordination between content experts and technology experts is especially critical (McSweeney, 2013); but programs can plan on the standard process of item development training, review and revision, and pilot and usability testing. Indeed, it is essential for demonstrating compliance with standards and best practices that AITs follow a parallel development process with traditional formats. Additional costs relate to additional tools, banking software, and time needed to increase the frequency with which items may need to be developed and rotated on/off test forms (discussed further in the Memorability/Test Security section). Furthermore, the format and sequence of items as well as usability testing will require additional development strategies for an assessment that includes AITs than one that includes only MCITs.

Moving from traditional MCITs, or from paper-based and performance-based item types to computer-automated assessments, can save time and other resources if the items can be scored programmatically. On the other hand, for programs moving to more complex item types—particularly those that require manual scoring—additional resources will need to be dedicated to developing scoring rubrics, recruiting and training (and possibly compensating) scorers, analyzing and monitoring results, and regulating rater behavior to ensure consistent standards and reasonable standard errors of measurement across items and examinees (Jones and Vickers, 2011).

Additional development effort also may be required for certain types of AITs. Consider a role-play simulation in which the response to one set of stimuli relays an examinee to a different set of follow up questions (also known as "branching" behavior). For example, consider whether answering A to Set 1 leads the examinee to Set 2, and whether answering B to Set 1 lead the examinee to Set 3. In this case, the path an examinee follows depends on how the examinee answers each set of questions and therefore the item bank must account for many possible outcomes. Some additional considerations for test developers creating branching items include determining whether to provide an opportunity for misdirected examinees to get back on the “correct” path and how to effectively analyze the seldom-selected path.

Finally, moving from traditional to computer-based items requires more integration between psychometricians or measurement professionals, test designers, developers, and technical staff (McSweeney, 2013). While the increased coordination and communication among these groups is yet another resource constraint, ultimately, those efforts will promote efficiency and likely yield the most enriched and efficient use of the data that comes with AITs.

On the other hand, most programs have multiple forms to reduce the exposure risk of “memorable” items from large candidate volumes and repeat examinees (Knapp, 2004). If item innovations are costly to develop, a program should consider the additional continued expense of maintaining multiple forms and rotating test content over time. On the other hand, if a program is already administering performance-based items that require expert judges to administer and/or score (e.g., portfolios, practical exams, role plays, analysis exercises), it can

be quite costly to gather examinees and judges in the same location at the same time. Eliminating the costs associated with travel, lodging, and compensation may offset the cost of the innovations. For example, the Bureau of Alcohol, Tobacco, Firearms, and Explosives successfully transitioned an assessment center for selection to computer delivery.[1] The ability to automate delivery and scoring of constructed responses (e.g., essays and short answer items) can increase the expedience and economy with which programs can administer performance tests.

To the extent that the test developer is able to create replicable item shells or templates (e.g., standard formats, programming, and coding) for item development, economies of scale can be leveraged, reducing the overall costs of developing and expanding the item bank (Downing, 2006; Muckle, 2012; Strain-Seymour et al., 2009). Templates provide parameters for item content to direct item writers' efforts and make paralleling and extending content easier.

> "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 online administration, making possible the goal of including alternative items in operational assessments in an efficient and sustainable manner." (Strain-Seymour et al., p. 8)

For example, programs using traditional MCITs typically have an item template that provides fields for item stem, response options, answer key, test blueprint area, and references. A program might similarly create and use a template for AITs (Parshall and Harmes, 2007; Strain-Seymour et al. 2009). Using the phlebotomy example again, a drag-and-drop item could be replicated to require the examinee to show the area from which s/he would draw blood from an arm, from a hand, etc. Another benefit of templates is that once item writers become familiar and comfortable with using them, they may be able to continue creating items with minimal assistance from software programmers (Strain-Seymour et al., 2009).

Another cost-related issue that should be considered in the design phase is interoperability, which is the ability of software or systems to communicate and exchange data across different storage and delivery platforms. Technical specifications exist that detail how item and assessment data are represented to allow for this exchange of information between systems, including the Question and Test Interoperability (QTI) specifications and the Accessible Portable Item Protocol (APIP) (IMS Global Learning Consortium, 2012; this information has been retrieved from http://www.imsglobal.org/question/qtiv2p1/imsqti_oviewv2p1.html). APIP also addresses accessibility needs that some examinees may require. The benefits of developing AITs in accordance with the QTI specifications include security of investment because the items or assessments will be portable to another vendor and possibly reduce time-to-market for custom item types or customizations to an existing item type because of the availability of ready-to-use templates (IMS Global Learning Consortium, October 2010). If custom item types are developed in a proprietary system, it may be difficult and costly for an organization to switch

[1] http://www.siop.org/UserFiles/Image/Refresh/Press-Release-Winners-FINAL.pdf#!

to a different vendor at some point in the future. More information regarding the QTI and APIP specifications can be obtained from the IMS Global Learning Consortium.

Scoring Alternative Item Types

Generally speaking, item scoring models can be classified into two broad categories: dichotomous (i.e., right/wrong) and polytomous (i.e., partial credit) (Parshall & Harmes, 2007). The number and type of available scoring models within each category are dependent on the specific characteristics of the item type. Traditional MCITs, for example, are typically scored dichotomously; although some testing programs) have used a formula scoring approach that penalizes examinees for incorrect responses as a way to remove the effects of guessing on overall test scores.

One appealing characteristic of AITs is the potential for obtaining richer information regarding an examinee's knowledge and/or skills. As a result, partial credit scoring models, which allow for finer levels of measurement, are regularly used with these items. For example, with a typical drag-and-drop item (i.e., matching), a program may potentially implement a partial credit scoring model that awards partial credit for each response option that is correctly matched to its target. A dichotomous scoring model, on the other hand, might only award full credit to an examinee who correctly matches all the response options to their targets and no credit for an examinee who correctly matches none or some of the response options to their targets. In general, for items that require more than one response or action, it is often a reasonable scoring strategy to award some level of partial credit for each response or action that the examinee performs successfully.

For some item types, guessing and other potential test-taking strategies may need to be considered when selecting a scoring model. For example, consider a multiple-choice multiple response item with four total response options, two of which are correct and two of which are incorrect. For this item, an examinee would likely be tasked with identifying the two correct options. If the scoring model specifies that an examinee is awarded one-fourth of a point for each of the four response options that is either correctly selected or correctly avoided, the examinee would then receive half credit (0.5 points) for either selecting all responses or skipping the question altogether (i.e., making no selections). For most testing programs, it would be unacceptable to award much, if any, credit to an examinee who simply followed a useful test-taking strategy, as opposed to having the knowledge necessary to earn that credit.

Partial credit scoring models may present other operational, technical, and/or logistical challenges as well. Most methods for establishing the passing score for an exam (i.e., standard setting) were developed with dichotomously scored MCITs in mind. Standard setting is already considered to be a cognitively complex task for the SMEs involved in the process, so incorporating AITs and partial credit scoring models adds to that complexity (and may potentially reduce the validity of the passing score). In addition to standard setting challenges, programs that use IRT need to obtain stable IRT item statistics for examinee scores to be accurate. Dichotomous IRT models (i.e., IRT models designed for items that are scored dichotomously) require smaller sample sizes than polytomous (i.e., partial credit) IRT models; however, only programs with sufficiently large candidate volumes would be able to support the use of a partial credit IRT-based scoring model. Furthermore, were a program that employs a computer adaptive testing (CAT) approach to incorporate AITs with partial credit scoring, the technical complexity involved in implementing that system would be daunting. Equally important,

explaining such a system to examinees in a way that was both accurate and understandable would likely be difficult.

The use of IRT requires its own special considerations when certain AITs are incorporated into an examination. Most testing programs use a "unidimensional" IRT model that assumes the test measures a single, overarching construct (e.g., accounting knowledge, nursing knowledge/skill). In order for this model to function as intended, however, all items must be independent of one another, meaning one's ability to respond correctly to one item must not depend on one's ability to answer another. Thus, in a simulation-type item that involved an opening scenario followed by a series of items tied to that scenario, the independence requirement would clearly be violated for the items within the series if examinees were unable to answer some items in the series correctly without correctly answering all (or some) of the previous items in the series. In this case, it might be necessary to treat the entire scenario as a single polytomous item with multiple components (i.e., the individual items in the series) that would all contribute to the total score for that item (i.e., the scenario).

In summary, most scoring-related challenges faced by a credentialing program when incorporating AITs (e.g., increased complexity in standard setting) can be overcome. But those challenges should not be ignored, and depending on the programs' specific situation, the resources required to overcome those types of challenges should be weighed against the benefits of including these item types in the examination.

ADA compliance/universal design considerations

Another consideration in selecting and implementing AITs relates to accessibility and compliance with the Americans with Disabilities Act (ADA). Use of AITs, particularly computer-based innovations such as drag-and-drop, graphics, and videos, must be designed to maximize accessibility for diverse examinees with a wide array of needs or to be adapted for accommodation requirements. Universal design (UD) principles are intended to be applied from the earliest stages of test design to eliminate distractions and irrelevancies.

All programs want items to discriminate examinees based on their ability on the designated construct of measurement. As described earlier, an advantage of AITs is the way in which they can reduce cognitive load. However, there may be other characteristics that disadvantage examinees who understand the construct being tested but may have difficulty with elements of the design.

Test development and delivery organizations have suggested universal design (UD) guidelines for developing and delivering items for computer-based tests. The primary intent of the UD guidelines is to ensure that the target construct is being measured as intended (Dolan, Burling, Rose, Beck, Murray, Strangman, Jude, Harms, Way, Hanna, Nichols, & Strain-Seymour, 2010). The UD guidelines provide a framework for identifying and organizing sources of variance associated with various item components, thus allowing for the reduction of construct irrelevant variance. A large part of the development framework is assessing the cognitive processes utilized to read, interpret, and respond to an item. Understanding how examinees with various disabilities process information differently can assist in reducing the potential for construct irrelevant measurement.

Just as advances in technology contribute to item innovations, such advances have also contributed to universal design principles. In addition to traditional types of accommodations

(e.g., readers, extra time, and large font), paper-based testing is now a form of accommodation. Computer-based test delivery companies provide private rooms and equipment adaptations (e.g., headsets, voice recognition technology, and screen reader software); furthermore, testing software complies with the various testing standards and state and federal regulations.

Usability and Pilot Testing

Regardless of item type or format of delivery, items should be pilot tested. AITs, particularly those that are technologically-enhanced, present new challenges. As such, it is strongly recommended that these items go through usability testing before they reach the pilot testing stage. Usability testing is most effective when incorporated into the item template design process described above (Parshall & Harmes, 2009). This is an iterative process in which item writing materials and prototype (or sample) items are developed, usability testing is conducted, and stakeholder review is performed. The data and information gathered by the usability testing and stakeholder review are used to refine and improve the design template. Usability testing has the added benefit of being closely tied with, and can inform decisions made about, ADA compliance and universal design.

Like any assessment, once the items have been developed, they must be pilot tested in settings identical to their intended operational settings and on subjects that represent the target population (Strain-Seymour et al., 2009). It is particularly important for a program that is transitioning to a new format to plan carefully and pay close attention to examinee perceptions, item performance, and item/test completion time.

The new items may take longer to answer. If so, a relevant question is whether there is adequate time in the test duration for examinees to be able to provide responses. For example, when a large information technology (IT) company moved one of its certification exams from task-based items to project-based performance items, it discovered during the field test that examinees spent much more time reviewing each data point and rechecking instructions – to the extent that many of them timed out of the test. Since the test was not designed to be completed quickly, the IT company had to revisit the timing (McSweeney, 2013).

Most certification tests are not speed dependent, so test time must account for any additional time examinees need. Significant questions are whether there is any additional value for the examinee and for the program if an increase in time is required, and can better results be obtained that can be balanced against increased expenses for seat time.

Memorability/Test Security Issues

Some authors have expressed a concern that AITs may be more capable of being remembered than traditional MCITs or text-based items (Knapp, 2004; Muckle, 2012; Sireci & Zenisky, 2006) because they are different from MCITs or contain media such as graphics, photos, or videos. This concern stems from research findings suggesting that unfamiliar or novel material are more easily remembered than familiar material (McDaniel, Dunay, Lyman, & Kerwin, 1988; Tulvig & Kroll, 1995; Waddill & McDaniel, 1998 as cited by Harmes & Wendt, 2009). However, this claim is not well supported by research conducted to date. A study by Harmes and Wendt (2009) found examinees did remember elements of AITs, such as how they interacted with the item (i.e., item format) and general content, but examinees did not generally remember enough specific content or keys that would compromise the items. Assuming drift in item difficulty could be related to item memorability, research conducted by Woo, Kim, and Qian (2014) also found

that AITs may not be more easily remembered than traditional MCITs. Woo, Kim, and Qian (2014) found that traditional MCITs got significantly easier over time, multiple response items got significantly more difficult over time, and fill-in-the-blank calculation items and ordered response items showed no significant drift in difficulty over time.

Whether or not the memorability of AITs poses a threat to test security, it is a testing industry best practice to ensure there are a sufficient number of items and exam forms and processes in place (e.g., retake limitations and waiting periods) to limit the overexposure of test items. One strategy for increasing the pool of AITs and reducing item exposure is to perform item cloning and create parallel forms (Sireci & Zenisky, 2006; Harmes & Wendt, 2009). During item cloning, variants of items are created by changing distractors, keys, or details in the item stem.

Examinee Preparation

If an examinee is unsure of how to evaluate and respond to an AIT, the item format has introduced construct irrelevant variance. Because the goal for all testing programs is to maximize the validity of the inferences made about examinee ability based on test scores, it is critical that examinees understand the function of AITs and how to respond correctly to them, given the examinee's job-related knowledge and skills. The importance of ensuring examinees are familiar with the item format and able to correctly respond to all item types on an exam is discussed in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014).

Crocker (2006) provides several suggestions for preparing examinees that are important to any program offering AITs, including pre-testing information (e.g., providing sample items with solutions and explanations of the problem-solving process, test day instructions that explain how to respond to item formats, and practice opportunities). Sireci and Zenisky (2006) recommend that examinees be provided with access to electronic sample exams (e.g., downloadable programs) so they can practice on high fidelity simulations; additionally, they recommend that a comprehensive tutorial be provided during the test session, prior to examinees beginning the test itself.

Test-day tutorials are critical, but may be insufficient (McSweeney, 2013). When a large IT organization introduced its performance-based certification exam, many of the examinees, familiar with the old system, skipped the tutorial and launched right into the exam. As a result, they were disoriented when they began taking the test. Fortunately, the organization learned this during the pilot test and was able to adapt the system. As important as in-test tutorials are, providing a tutorial opportunity before the test administration may be even more advisable, especially if the use of AITs represents a marked departure from previous testing experiences. Examinees should have an opportunity to practice sample questions in the new formats in a relaxed environment, unaffected by the pressures and anxieties of the actual testing occasion. Tutorials external to the test itself give examinees an opportunity to confront and resolve any unfamiliarity with the novel formats, which otherwise may have an undesirable impact on their performance on the test. While this type of tutorial may involve additional cost, it is critical to give non-"tech-savvy" examinees an opportunity to practice with the new item types before the live test event.

Examinee Reactions and Considerations

The available data gathered regarding examinee reactions to AITs has primarily been gathered from post-test surveys, but some anecdotal feedback has also been documented in available literature. While the survey questions, associated assessments containing AITs, and examinees were different in each instance, there seem to be some trends across the available literature. Most examinees reported: (a) a generally positive reaction to assessments containing AITs (Baumann, Steinmetzer, Karami, & Shafer, 2009; Muckle, 2012; Strain-Seymour, Way, & Dolan, 2009); (b) few issues in understanding how to respond to the AITs (Dolan, Goodman, Strain-Seymour, & Sethuraman, 2011; Muckle, 2012; Wendt and Harmes, 2009a); (c) the AITs were more difficult than traditional MCITs (Dolan, Goodman, Strain-Seymour, & Sethuraman, 2011; Wendt & Harmes, 2009b); and (4) the AITs were more engaging and realistic than MCITs (Dolan, Goodman, Strain-Seymour, & Sethuraman, 2011; Strain-Seymour, Way, & Dolan, 2009, Wendt & Harmes, 2009a).

Dolan, Goodman, Strain-Seymour, Adams, and Sethuraman (2011) conducted a cognitive laboratory study to assess a student's cognitive processing steps when responding to an AIT, the degree these steps correspond with expected steps, and the degree that enhanced functionality of AITs impact a student's responses. During the experiment, 36 students logged onto a web conference where they were asked to respond to the items and verbalize their thought process as they did so. The results of the study found that students experienced very few usability issues when responding to individual items, including items with more complex interfaces. Students with greater computer experience tended to respond to the items faster than less experienced computer users, but they were not more likely to get the item correct. They also found that the AITs allowed students to take multiple paths to arrive at their final response, students were highly engaged in the task (even in this low stakes environment), and students' steps and missteps corresponded with the expected steps for the items.

As part of ongoing research on AITs, Strain-Seymour, Way, and Dolan (2009) have collected anecdotal feedback from students, educators or teachers, curriculum designers, and content experts. They note the feedback was generally positive and seemed to be consistent with available research findings. Student feedback gathered in a text box at the end of examinations administered in relation to a state testing program indicated that students seemed to enjoy the level of interactivity of AITs, the continuity from the classroom or lab experiences, and the helpful visualizations. The researchers also noted some trends in feedback that were received when curriculum specialists, teachers, and content area experts reviewed the AITs during the development phases. Curriculum developers had a positive response to AITs testing core multistep processes (e.g., multi-step real life mathematical problems) that can test granularity in the process without removing context or complexity. Teachers and educators were enthusiastic about the high-level of continuity between the classroom or lab activities and the AITs. Teachers and educators also tended to have a more positive response to the transition to technology-based assessments when AITs are involved.

<u>Best Practices for Examinee Considerations</u>

Based on the above research on examinee reactions to AITs, ATP and ICE recommend that credentialing programs adopt the following practices and/or procedures to increase the probability that examinees will have a positive experience with AITs.

- Communicate the changes to the test format to examinees and stakeholders in advance of implementing the changes. It may also be helpful to communicate the changes in as many venues a possible (e.g., website, conference or other meetings, newsletter, candidate information bulletin, or other communications with candidates) so that examinees are not taken by surprise on examination day (Muckle, 2012).

- Provide sample items or tests, or interactive tutorials mimicking the functionality of the new formats, so that examinees can ensure they understand the structure and approach of the AITs and can become familiar with how to respond to the new items types and practice interacting with them. This can significantly reduce the examinee's level of anxiety (Strain-Seymour, Larkin, & Goodman, 2011). Providing sample items may be especially helpful since more than one study indicated that examinees either did not read the details of the instructions at first or needed some assistance in figuring out how to respond to the items.
- Evaluate the amount of time that should be provided to examinees to complete the examination when adding AITs since more than one study found that AITs were more time consuming than traditional MCITs. If including AITs in an assessment requires an increase in the examination duration, ensure this is also clearly communicated to examinees in advance of the exam.
- Ensure that the instructions for the AITs have been pilot tested with the items to ensure that they are clear and easily understood by examinees.
- Explain to examinees how each type of AIT, along with the overall assessment, will be scored,

Summary and Concluding Remarks

A credentialing program that is considering including AITs in its assessment(s) needs to complete a full analysis and evaluation to determine whether AITs would add value to its certification program. If so, the organization then needs to ensure it understands the associated cost and resource requirements to properly develop and implement AITs in the assessment(s). Without this initial groundwork, the program could actually reduce the quality of measurement of its assessments(s) and/or make inefficient investments in item types that do not add value to its program.

Part of this initial groundwork is weighing the benefits and costs associated with developing and maintaining AITs for its assessment(s). The various considerations have been discussed in detail in the paper, but summarized below are five steps a credentialing organization can use in its feasibility evaluation process.

Step 1: Assess the current certification program to identify the constructs that are being measured well and constructs that are measured inadequately or not at all by existing item types constructs. This assessment should be made in relation to current job/role/practice analysis data to ensure all knowledge and skills measured by the assessment are currently important to competent job performance in relation to the scope and level of the credential. If the certification program does not have a recent job/role/practice analysis study, then it will be critical to conduct this study before proceeding with the AIT evaluation process.

If the certification program has a recent job/role/practice analysis and it is determined that critical constructs in the job/role/practice analysis are currently being measured inadequately or not at all, the credentialing organization should evaluate if AITs could improve or expand the current measurement capabilities. If so, the credentialing organization should also make an initial determination regarding the optimal scoring method (dichotomous, partial credit, etc.) for each AIT. The AITs and desired scoring methodologies should be determined without considering the item types and scoring options currently available in vendor software (Muckle, 2012; Parshall & Becker, 2008; Becker, 2010). SMEs, psychometricians, stakeholders,

managers, examinees, and test users should all be involved in the evaluation and design steps of this endeavor (Bontempo, 2001).

Unfortunately, to date, there is still not much research on which item types are most effective at measuring various types of content. Available research should be reviewed and, when research is not available, the credentialing organization should consult with qualified psychometricians, assessment development professionals, SMEs, and other stakeholders to provide documentation that the chosen AITs are appropriate ways to assess an examinee's level of competence on identified domains or constructs. If feasible, confirmatory research (or post-implementation evaluation) should be planned for after the AITs have been developed and implemented to ensure there is evidence of construct validity and not just face validity (i.e., the perception that items are measuring job-related knowledge and skills).

Step 2: Once AITs and preferred scoring methodologies are identified, the credentialing organization should develop a clear plan for the development and maintenance of these items and design some initial templates for the AITs. As mentioned above, the organization should include SMEs, assessment development professionals, psychometricians, and other stakeholders in this planning and design phase. The plan for development and maintenance should be as detailed as possible in order to complete the next steps of determining the costs and resource requirements. Parshall and Harmes (2009) provide a model for the design and development of AITs that might be useful to identify required steps. Compared to the development of traditional MCITs, the development of AITs frequently requires additional steps (e.g., usability or user acceptance testing, more complex scoring and statistical analysis of items).

Step 3: Once a development plan has been created and initial templates of the AITs have been developed, the credentialing organization should determine if available test delivery software has the capability of administering and scoring the AITs designed for its assessment(s) or if additional software development will be required. If software development will be required, the organization should obtain cost estimates before proceeding because the costs can greatly vary. If the credentialing organization plans to develop a custom item type, it should also consider interoperability factors.

Step 4: The credentialing organization should also assess the costs and resource requirements to execute the AIT development plan. The cost estimate should include any external resources required (e.g., psychometrician, software developer/programmer), internal staff time to coordinate project and SMEs, and travel costs if in-person meetings are utilized. Be aware that the development of AITs is often an iterative process in which the templates go through multiple rounds of usability testing and revision. Additionally, item writing guidelines and training materials need to be developed specific to these item types for the item writers. If the items require any manual scoring, costs and resources for rater training and scoring procedures should also be evaluated.

Step 5: To ensure a successful implementation of AITs in its assessment(s), a credentialing organization should also develop a plan and assess the associated costs of communicating the change to stakeholders. This may include the development of sample items or tests so examinees could practice interacting with the AITs before the test day.

Transitioning away from a traditional item or delivery format may seem like a major paradigm shift for a testing organization. However, the foundation on which any test stands is its reliability and validity, and those fundamental qualities are independent of the tools used to measure a

specific domain or set of domains. As technology continues to change and the science of psychological measurement continues to mature, AITs and delivery formats will undoubtedly continue to be developed and researched. Therefore, the guiding philosophy for testing organizations exploring or implementing AITs should be to ensure the reliability of the scores and the validity of the inferences.

References

AERA, APA, & NCME (2014). *Standards for educational and psychological testing*. Washington, D.C.: Author.

Baumann, M., Steinmetzer, J., Karami, M. & Shafer, G. (2009). Innovative electronic exams with voice in- and output questions in medical terminology on a high taxonomic level. *Medical Teacher 31*, e460-e463. doi:10.3109/01421590902842433

Becker, K. (2010). The care and feeding of innovative items. Presented at the 2010 Annual Meeting of the National Council on Measurement in Education, Denver, CO.

Bontempo, B. (2001). Innovative Item Types: How to Evaluate the Bang for the Buck. Paper presented at the Computerized Testing Conference sponsored by the Association of Test Publishers, Tucson, AZ.

Collins, M. & Waugh, G.W. (November, 2008). *Beyond Multiple Choice: The Development and Implementation of Alternative Item Types in a Certification Exam*. Poster presented at the annual meeting of the National Organization for Competency Assurance, San Francisco.

Collins, M., Keenan, P., & Ramli, M. (2008). *Development and implementation of the Ratings Veterans Service Representative (RVSR) skills certification test* (FR-09-05). Alexandria, VA: Human Resources.

Crocker, L. (2006). Preparing examinees for test taking: Guidelines for test developers and test users. In S.M. Downing & T.M. Haladyna (Eds.). *Handbook of Test Development* (pp. 115-128). Mahwah, NJ: Lawrence Erlbaum Associates.

Dolan, R.P., Burling, K.S., Rose, D., Beck, R., Murray, E., Strangman, N., Jude, J., Harms, M., Way, W.D., Hanna, E., Nichols, A., & Strain-Seymour, E. (2010). Universal design for computer-based testing. Bloomington, MN: Pearson.

Dolan, R.P., Goodman, J., Strain-Seymour, E., Adams, J., & Sethuraman, S. (2011, March). *Cognitive Lab Evaluation of Innovative Items in Mathematics and English Language Arts Assessment of Elementary, Middle, and High School Students.* Retrieved July 1, 2013, from http://www.pearsonassessments.com/hai/images/tmrs/Cognitive_Lab_Evaluation_of_Innovative_Items.pdf.

Downing, S.M, Baranowski, R.A., Grosso, L.J., & Norcini, J.J. (1995). Item type and cognitive ability measured: The validity evidence for multiple true-false items in medical specialty certification. *Applied Measurement in Education, 8*(2), pp. 187-197.

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

Dwyer, A.C., Penny, J. A. & Johnson, R.L. (2015). Scoring alternative item types: There's many a slip between the cup and the lip. Paper presented at the Association of Test Publishers Innovations in Testing Conference, Palm Springs, CA.

Harmes, J.C., & Wendt, A. (2009). Memorability of innovative items. *Clear Exam Review*, *20*(2), 16-20.

Huff, K.L. & Sireci, S. G. (2001). Validity Issues in Computer-Based Testing. *Educational Measurement: Issues and Practice, 20,* 16-25.

IMS Global Learning Consortium, Inc. (2012, August 31). *IMS question & test interoperability: An overview (final specification version 2.1)*. Burlington, MA: Author. Retrieved May 31, 2016 from http://www.imsglobal.org/question/qtiv2p1/imsqti_oviewv2p1.html.

IMS Global Learning Consortium, Inc. (2010, October). *IMS Question & Test Interoperability Specification: A Review*. Burlington, MA: Author.

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

Jones, M. & Vickers, D. (2011). Considerations for performance scoring when designing and developing next generation assessments. (White Paper.) Bloomington, MN: Pearson.

Knapp, D.J. (2004). A discussion of innovative measurement methods. Presentation for HumRRO, Alexandria, VA.

Krogh, M.A. & Muckle, T. J. (2017). Assessing the psychometric properties of alternative items for certification. *Journal of Applied Measurement*, 17.

McBride, J. R., & Martin, J. T. (1983). Reliability and validity of adaptive ability tests in a military setting. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 223-236). New York: Academic Press

McDaniel, M.A., Dunay, P.K., Lyman, B.J., & Kerwin, M.L.E. (1988). Effects of elaboration and relational distinctiveness on sentence memory. *American Journal of Psychology, 101*, 357-369.

McSweeney, S. H. (October 10, 2013). The Birth of a New Item Type. Webinar presented to the Performance Testing Council.

Muckle, T.M. (2012). Beyond multiple choice: Strategies for planning and implementing an innovative item initiative. Washington, DC: Institute for Credentialing Excellence.

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

Parshall, C.G., & Harmes, J.C. (2007). Designing templates based on a taxonomy of innovative items. Paper presented at the Graduate Management Admission Council Conference on Computerized Adaptive Testing in Minneapolis, MN.

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

Parshall, C.G., & Harmes, J.C. (2009). Improving quality of innovative item types: Four tasks for design and development. *Journal of Applied Testing Technology, 10*(1), pp. 1-20.

Parshall, C.G., Harmes, J. C., Davey, T., & Pashley, P.J. (2010). Innovative items for computerized testing. In van der Linden, W. J.; Glas, C. A. W. (Eds.) *Elements of Adaptive Testing*. (pp. 215-230). New York, NY: Springer.

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

Ramstad, R. (2013, October). *Model for Innovative Item Development.* Bloomington, MN: Pearson VUE.

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).

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

Strain-Seymour, E., Larkin, J., & Goodman, J. (2011). Innovative Items and the Challenges of Race to the Top. Paper presented at the Association of Test Publishers Innovations in Testing Conference in Phoenix, AZ.

Strain-Seymour, E., Way, W.D., & Dolan, R.P. (2009). *Strategies and Processes for Developing Innovative Items in Large-Scale Assessments*. Iowa City, IA: Pearson Education.

Tulving, E., & Kroll, N. (1995). Novelty assessment in the brain and long-term memory encoding. *Psychonomic Bulletin & Review, 2*, 387-390.

Waddill, P. J., & McDaniel, M.A. (1998). Distinctiveness effects in recall: Differential processing or privileged retrieval? *Memory & Cognition, 26*, 108-120.

Wan, L. & Henly, G.A. (2012). Measurement properties of two innovative item formats in a computer-based test. *Applied Measurement in Education*, 25(1), 58-78

Wendt, A., & Harmes, J.C. (2009a). Evaluating innovative items for the NCLEX part 1, Usability and pilot testing. *Nurse Educator, 34*(2), 56-59.

Wendt, A., & Harmes, J.C. (2009b). Developing and evaluating innovative items for the NCLEX part 2, Item characteristics and cognitive processing. *Nurse Educator, 34*(3), 109-113.

Woo, A., Kim, D., & Qian, H. (2014). Exploring the psychometric properties of innovative items in CAT. Paper presented at the 14th annual Maryland Assessment Conference, College Park, MD.

Appendix: Sample Alternative Item Types

Overview of Item Types

- **Multiple Choice**
  - MC with Audio/Video Prompt
  - MC with Graphics
  - Multiple Choice Multiple Response
  - Discrete Option Multiple Choice
  - Table Layout
  - Drop-Down Menu
- **Constructed Response**
  - Free Response / Essay
  - Fill In The Blank
  - Short Answer
  - Spoken Response
- **Hot Spot**
  - Single Response
  - Multiple Response
  - With Audio Prompt
  - Plotting
- **Drag & Drop**
  - Matching
  - Ranking/Ordering
- **Simulation**
  - Semi-Interactive Console
  - Interactive Spreadsheet
  - Interactive Line Chart
  - Code Simulation
  - Simulation with MCQs

Acknowledgement

The following organizations have generously supplied the item type examples included here. With their consent, their item types have been included. We thank them for the participation.

Donating organizations
American Board of Pediatrics
American Nurses Credentialing Center
American Registry for Diagnostic Medical Sonography (ARDMS)
Association for Financial Professionals
Caveon Test Security
Center for Educational Measurement, Excelsior College
Cisco Global Certifications
Commission on Dietetic Registration
Graduate Management Admission Council® (GMAC®)
Learnosity
MDCB
Mettl
Microsoft Learning Experiences
National Board of Certification and Recertification for Nurse Anesthetists (NBCRNA)
National Center for Competency Testing (NCCT)
Open Assessment Technologies S.A., publisher of the TAO open source assessment platform,
Pediatric Nursing Certification Board
Testrac.com, Ltd.
The American Registry of Radiologic Technologists

Disclaimer: Neither ATP nor ICE is endorsing any of the item types that are included in this portfolio. These are included for demonstration purposes only and test sponsors should work with a psychometrician and subject matter experts to determine if an item type will add measurement value to its assessment(s).

Innovative Multiple Choice

Multiple Choice with Audio Prompt

Multiple Choice with Video Prompt

How to Play Video Exhibits

There may be some items that include a video. When the Exhibit button is clicked, a window opens that displays the video. To play the video press the arrow (►) button. You will not be permitted to leave the item until you have viewed the entire video.

The item below uses a video exhibit. Practice opening, playing, and closing the video exhibit window.

Video Controls:

Play **Pause** **Stop** **Loop** (press play, then loop to play video continuously)

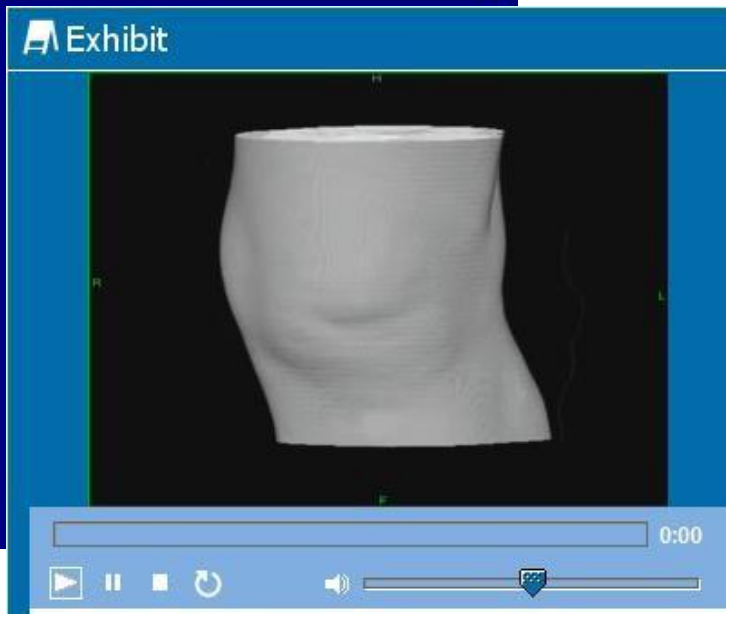
The correct answer is "B. knee." Select response "B" now and press the ***Next*** button to continue.

Click the ***Exhibit*** button to view the video.

Exhibit

What joint does the video demonstrate?

- A. shoulder
- B. knee
- C. ankle
- D. elbow



Multiple Choice with Video Prompt

Stem:

A 3.7-kg newborn infant is cyanotic immediately after birth. By 4 hours of age, cyanosis has increased and dyspnea and tachypnea develop. Percutaneous oxygen saturation is 55%. The echocardiogram is **SHOWN**.
Which of the following would be of most benefit to this infant?

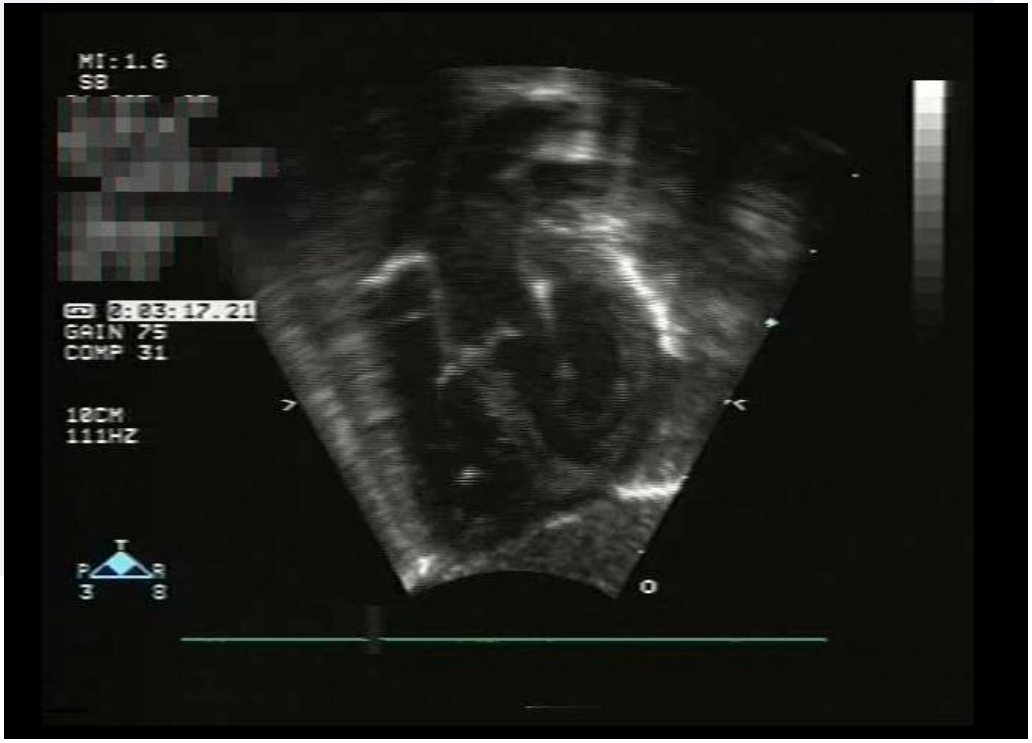
Assets:

CARD600228.mp4 : Click to Play Download File

Options:

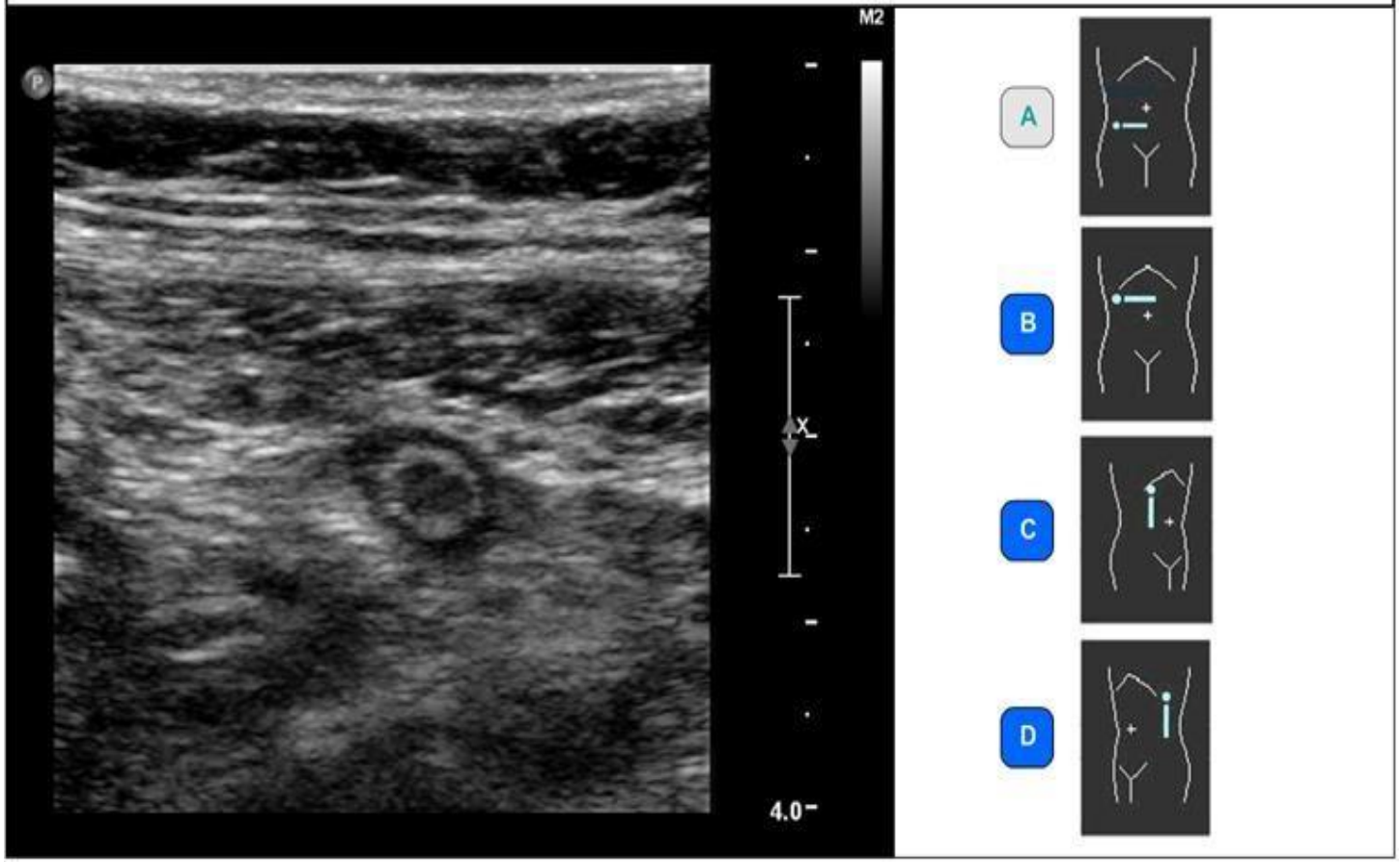
- [ ] A. Assisted ventilation
- [x] * B. Balloon atrial septostomy
- [ ] C. Nitric oxide therapy
- [ ] D. Pulmonary balloon valvuloplasty
- [ ] E. Sodium bicarbonate therapy

A screen shot of the video is below.



Multiple Choice – Graphics as Options

(1) Using your mouse, select which pictogram accurately demonstrates the location from which this abdominal image was obtained.



Multiple Response

Number Of Keyed Responses Identified

The patient has completed his treatment and is seen in a six-month survivorship visit. Much to the patient's disappointment, his weight remained stable during his treatment. The patient is motivated to make healthier lifestyle choices and asks the dietitian about what he can do to reduce the risk of recurrence. Which lifestyle recommendations would the dietitian offer to this patient? **Select three.**

A. Adopt strict vegetarian lifestyle
B. Attain and maintain a healthy weight
C. Incorporate 75 minutes of vigorous activity daily
D. Incorporate at least 2 servings of red wine daily for cardiac health
E. Increase intake of fruit and vegetables to at least 5 servings daily
F. Incorporate 150 minutes of moderate activity weekly
G. Maintain current weight

Multiple Response
Number Of Keyed Responses Not Identified

Which of these cities are state capitals?

- ☐ Wilmington, NC
- ☐ Trenton, NJ
- ☐ Topeka, KS
- ☐ St. Louis, MO

Multiple Response with Multiple Exhibits

One of your employees is unable to hear audio from his computer.

You review the information provided by the user in Support Report #1234567890, the computer's Device Manager (click the **Exhibit** button to view these documents), and Playback settings (displayed below).

Exhibit 1

Exhibit 2

Exhibit 3

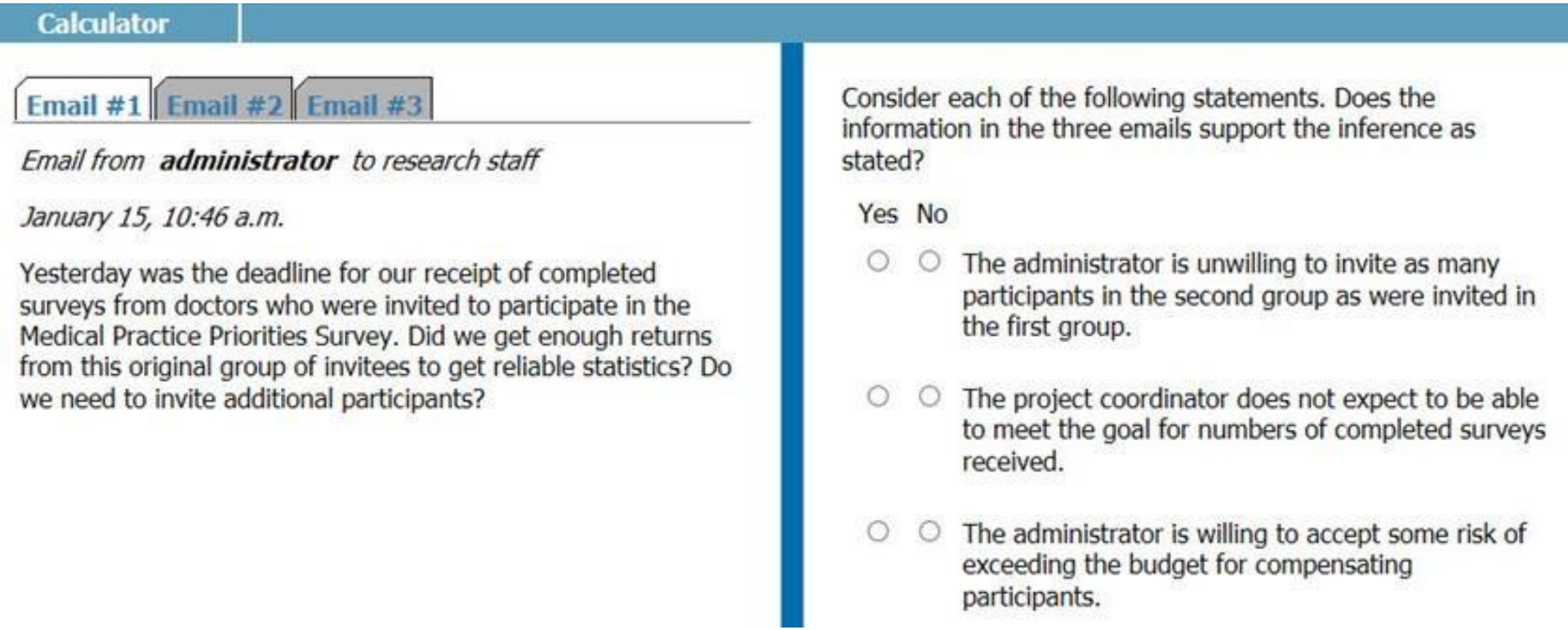
Support #1234567890

User reports he cannot hear sound through his computer.
User connects to a monitor with speakers by using an HDMI cable.
User does not know if the computer has an internal audio device.

Answer Area

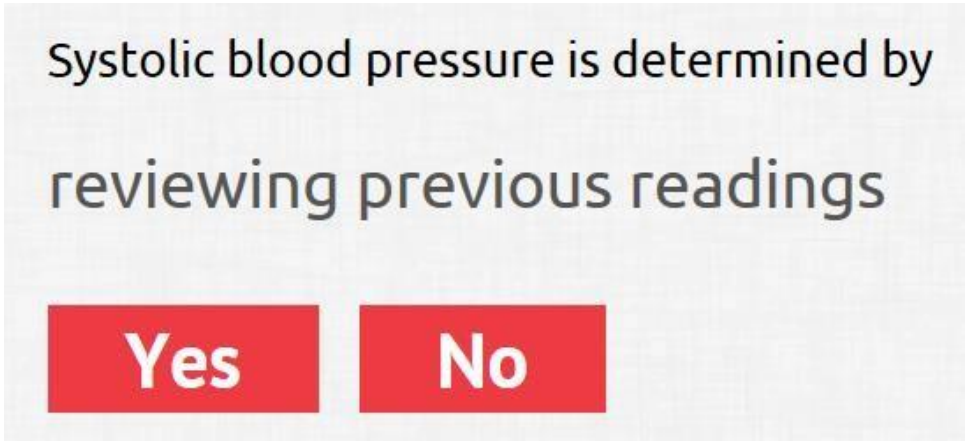
	Yes	No
The computer has a sound card installed and it is working properly.	○	○
The HDMI Device should be set as the default device.	○	○
The user should connect a separate audio cable from his computer sound card directly to the monitor.	○	○

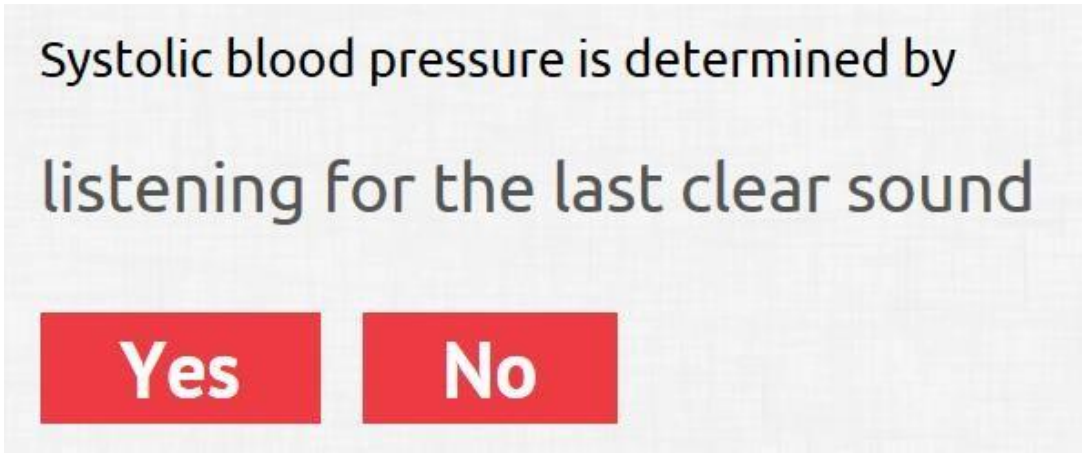
Multiple Response with Multiple Exhibits

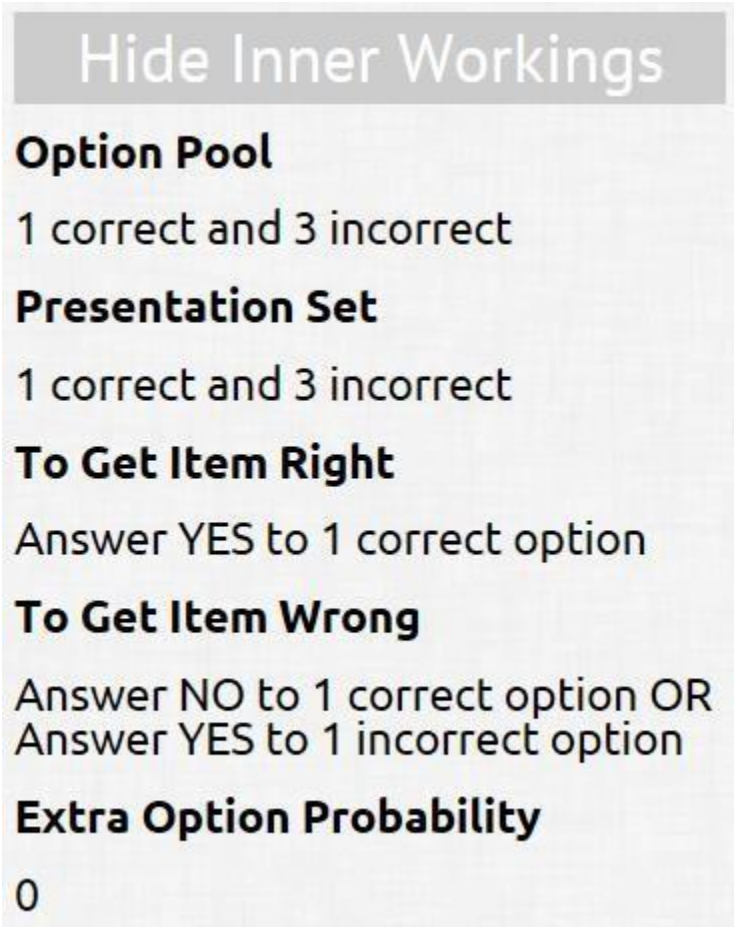


Test taker is provided multiple exhibits on the left side of the screen (in this case, three emails). Test taker reads emails and responds to each item on the right side of the screen.

Discrete Option MC







Test taker is presented the stem and one answer option at a time and must respond Yes or No to each answer option. The presentation and scoring settings on these items are customized by the test sponsor. On the same item, test takers may not be presented the same answer options or the same number of answer options.

Table Layout

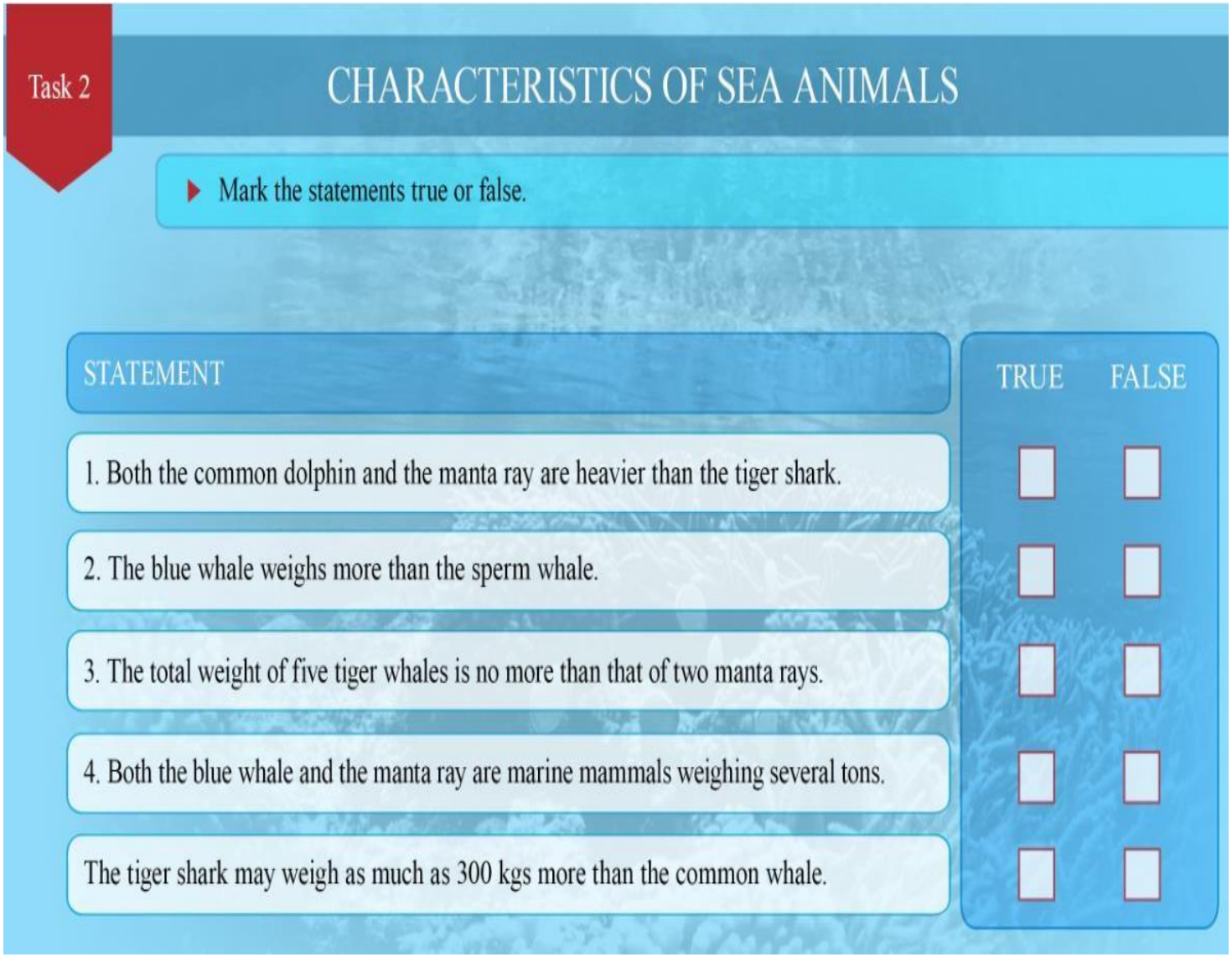


Table Layout (Two-Part Analysis)

Contoso is adding two locations, Paris and Amsterdam, to its network environment. Contoso's business requirements include the following:

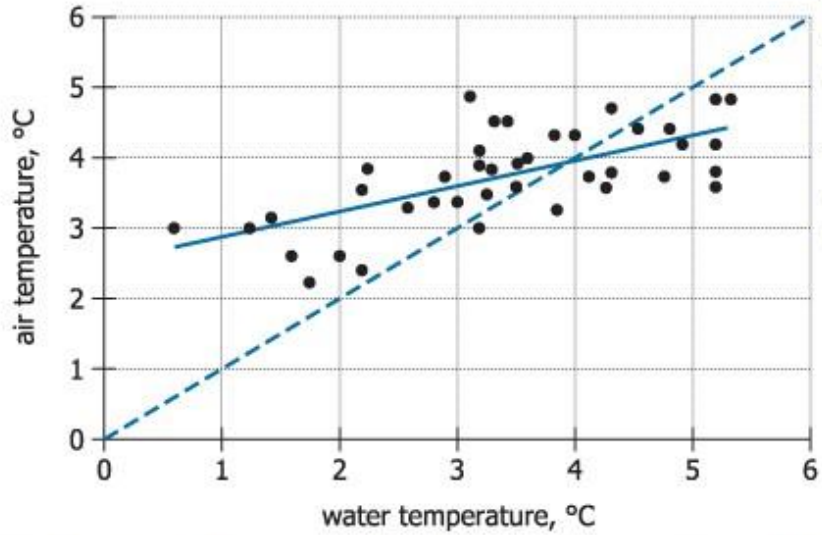
- All external access to the organization of the research.contoso.com domain is provided through the Internet link at the Paris office
- All inbound and outbound email for the domain goes through an email appliance in the Paris office.
- The hardware load balancer deployed to the Amsterdam office must bridge all SSL connections to the Exchange servers.

In the table below, identify the server role that must be placed in each location to ensure Contoso's business requirements are supported. Make only one selection in each column.

Answer Area

Server Role	Paris	Amsterdam
File Server	○	○
Global Catalog Server	○	○
Domain Controller	○	○
Schema Master	○	○

Drop-Down Menu Options



The graph at the left is a scatter plot with 40 points, each representing the temperature of the ocean water, measured at a fixed location off the coast of West Iceland, and the air temperature, measured on land at a fixed location in West Iceland. Both the water temperature and the air temperature, in degrees Celsius, were measured at noon on Wednesday of each of 40 consecutive weeks last year. The solid line is the regression line and the dashed line is the line through the points (0,0) and (6,6).

Use the drop-down menus to fill in the blanks in each of the following statements based on the information given by the graph.

The relationship between the water temperature and the air temperature is Select...

The slope of the regression line is Select... the slope of the dashed line.

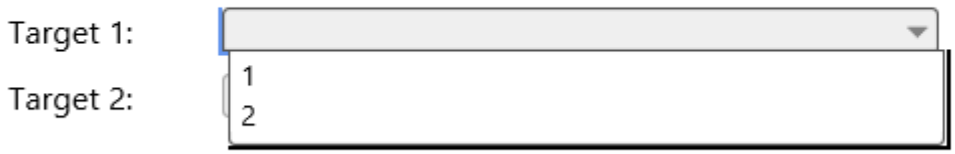
Select...
less than
greater than
equal to

Drop-Down Menu of Options

Please select an answer from the drop down list:

```
active screen item()
{
  select the = Correct 1
  if (color != red)
  {
  color.Blue 3
  }
}
```

What is the correct answer? (To answer, select the correct choice int the answer area.)



Constructed Response

Free Response / Essay

Fill-In-The-Blank

Fill in the blanks.

Sherlock Homes had sprung out and seized the _____ by the collar. The other dived down the hole, and I heard the sound of _____ cloth as Jones clutched at his skirts. The light flashed upon the barrel of a revolver, but Holmes' _____ came down on the man's wrist, and the pistol _____ upon the stone floor.

Fill-In-The-Blank with Drop-Down Menus

Fill in the blanks

"It's all clear,' he [ ]. "Have you the chisel and the bags? Great Scott! Jump, Archie, jump, and I'll swing for it!'

Sherlock [ ] had sprung out and seized the [ ] by the collar. The other dived down the hole, and I heard the sound of [ ] cloth as Jones clutched at his skirts. The light flashed upon the barrel of a revolver, but Holmes' [ ] came down on the man's wrist, and the pistol [ ] upon the stone floor.

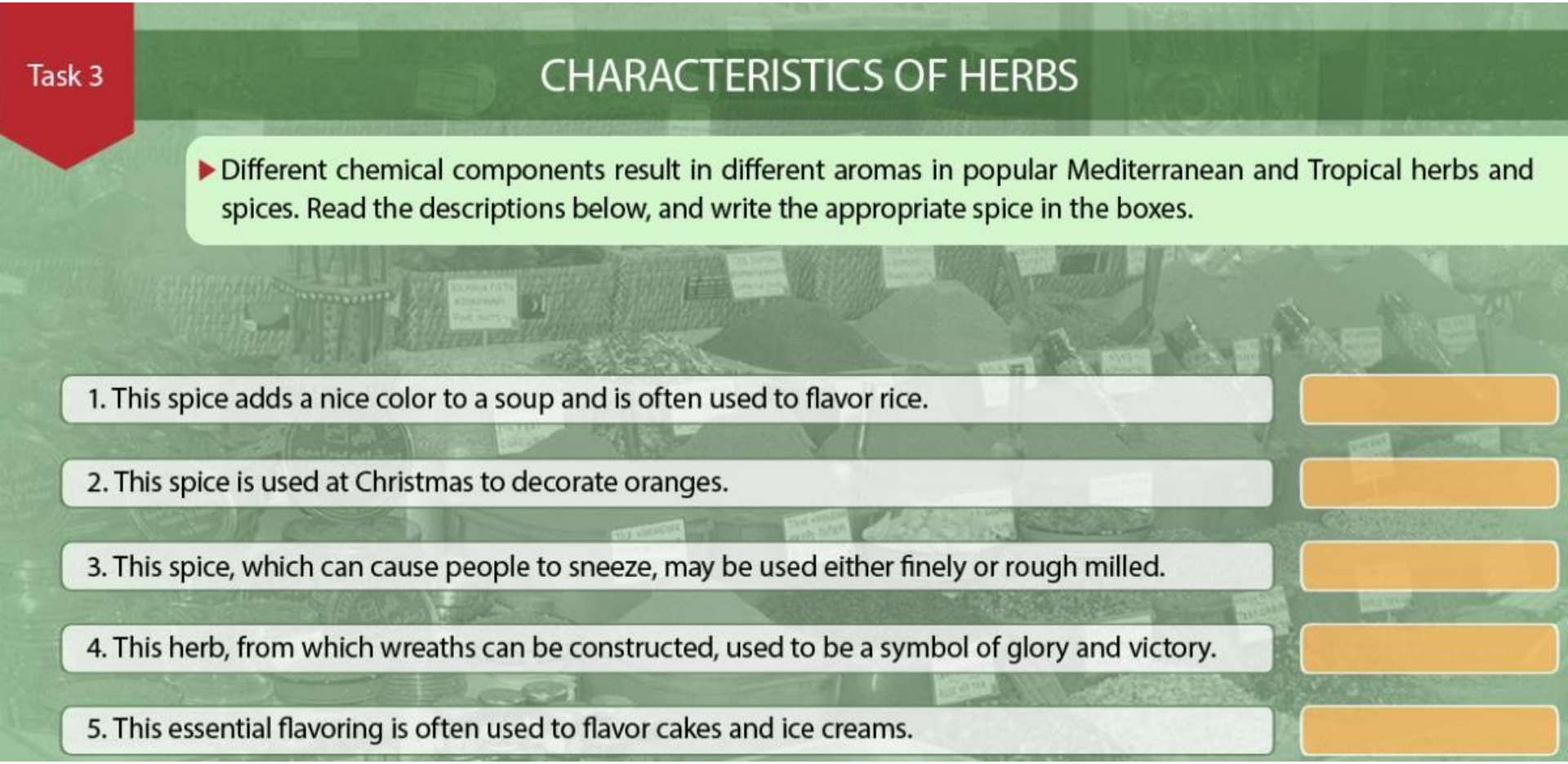
Short Answer

Who is the Mayor of New York City?

The item is scored as follows:

"Michael Bloomberg" gets one point, "Bloomberg" gets half a point.

Short Answer



Short Answer – Numeric

Calculate the cardiac output given the following hemodynamic parameters:

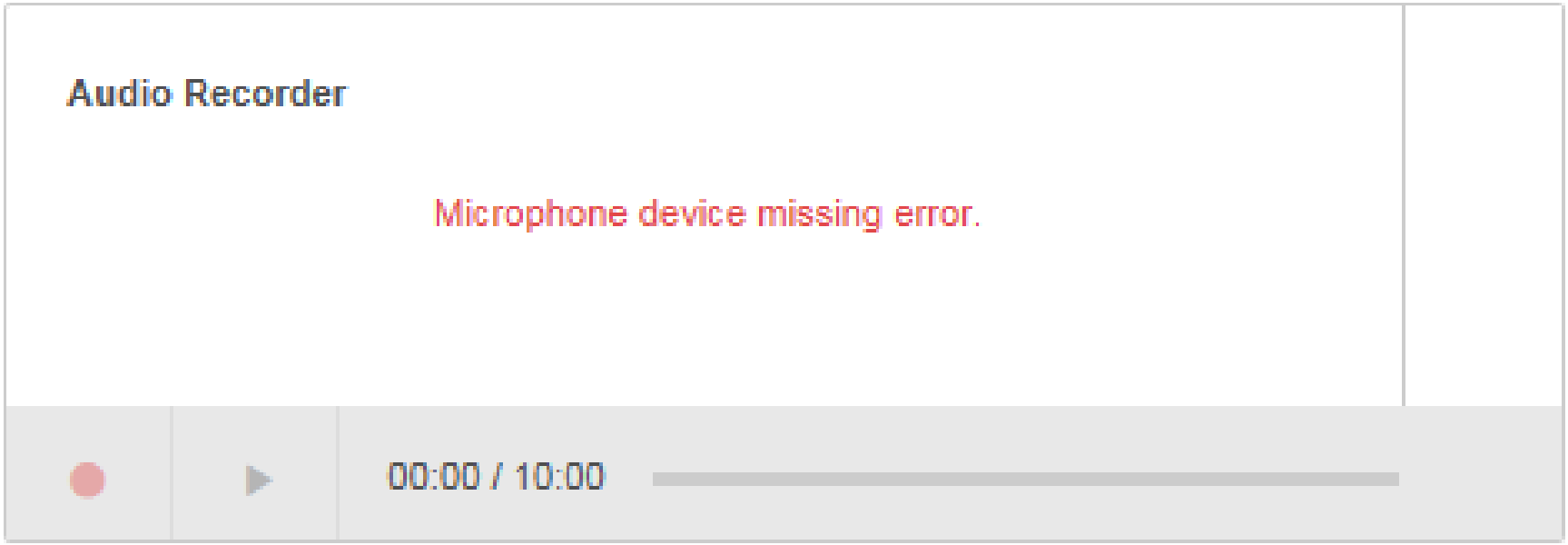
Stroke volume: 60 ml/beat; Blood pressure: 150/70 mmHg; Heart rate: 50 per min

Enter your answer below as a whole number (no decimals) in L/min.

L/min

Spoken Response

Describe a typical day in your life.

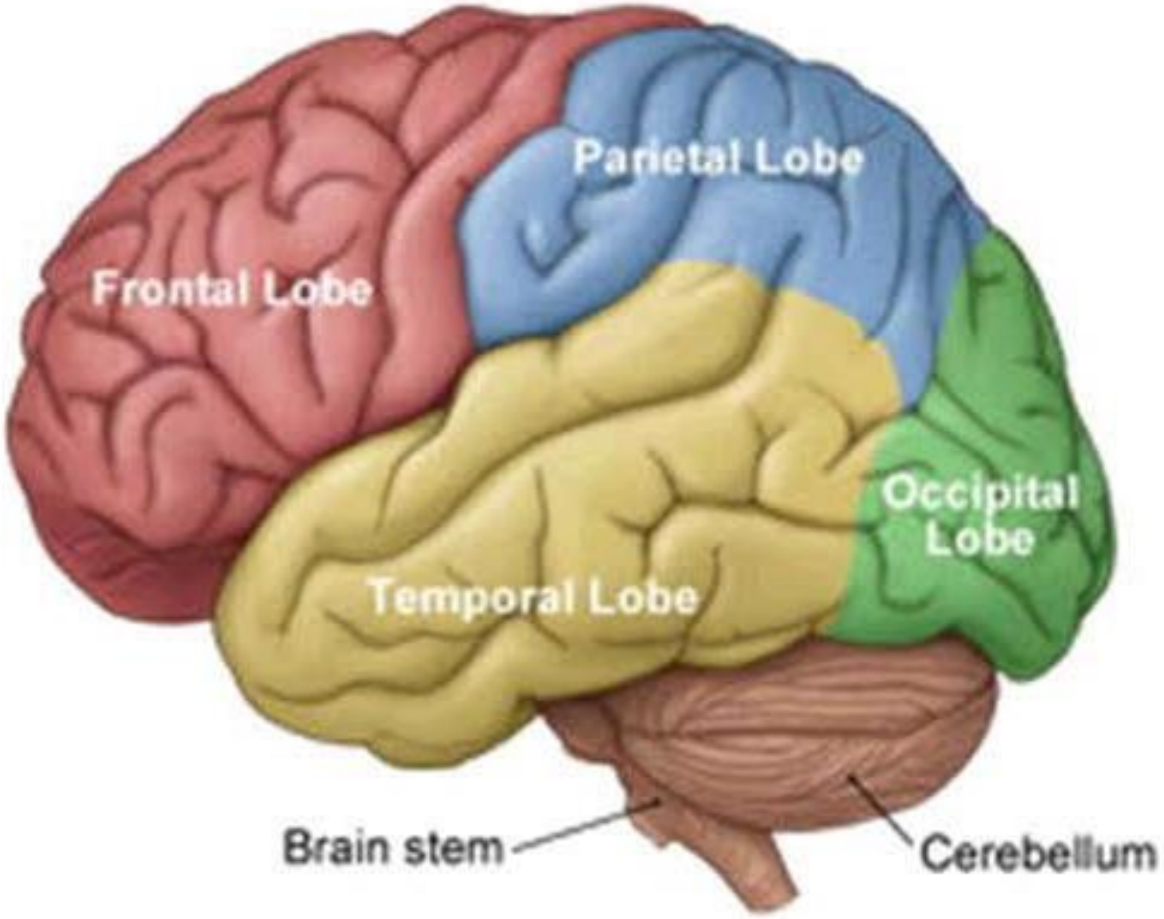


This item type is frequently used to test language proficiency or translation skills.

Hot Spot

Hot Spot – Single Response, Single Correct

Click on the region of the brain which is the primary visual reception area

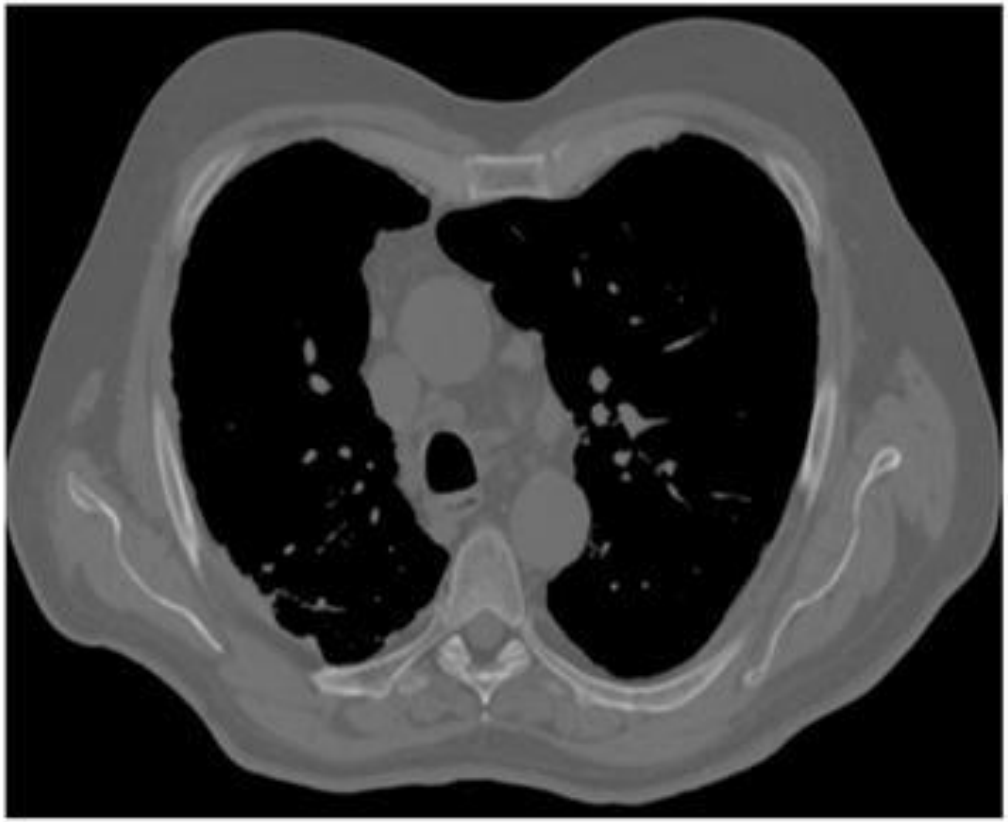


Hot Spot – Single Response, Single Correct

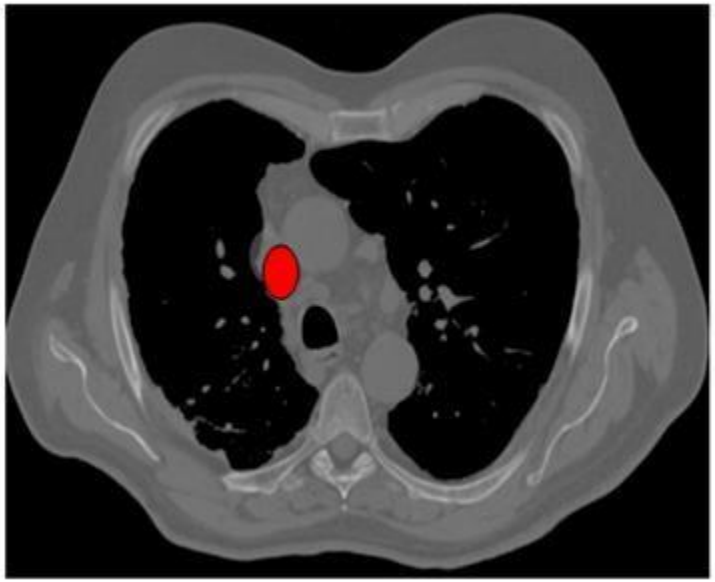
Click on the condition that an adult-gerontology primary care nurse practitioner appropriately treats with cryotherapty.

Hot Spot Single - Response, Single Correct

On the axial CT below, identify and click on the Superior Vena Cava with your Cursor:

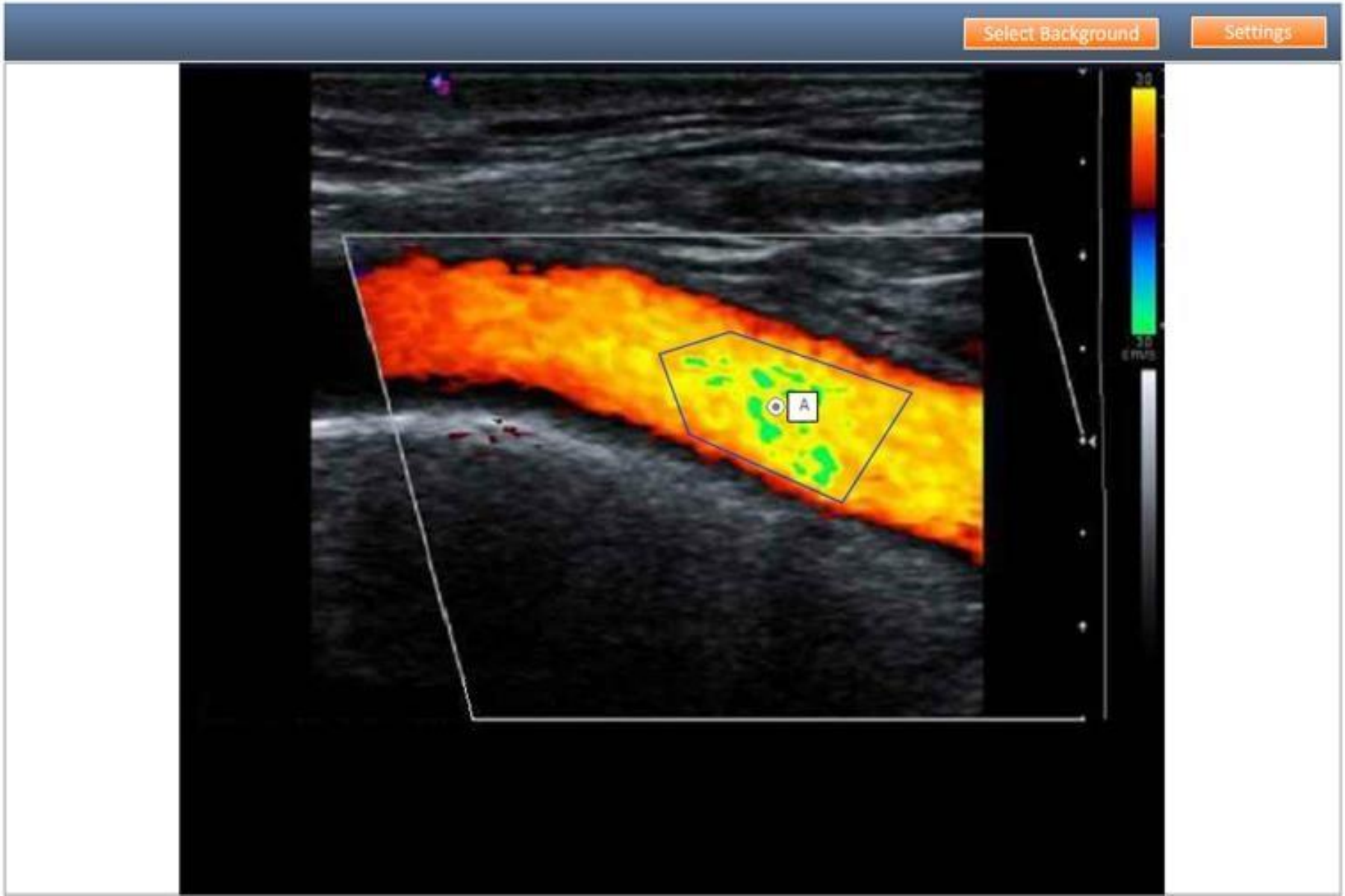


With answer selected



Hot Spot - Single Response, Single Correct

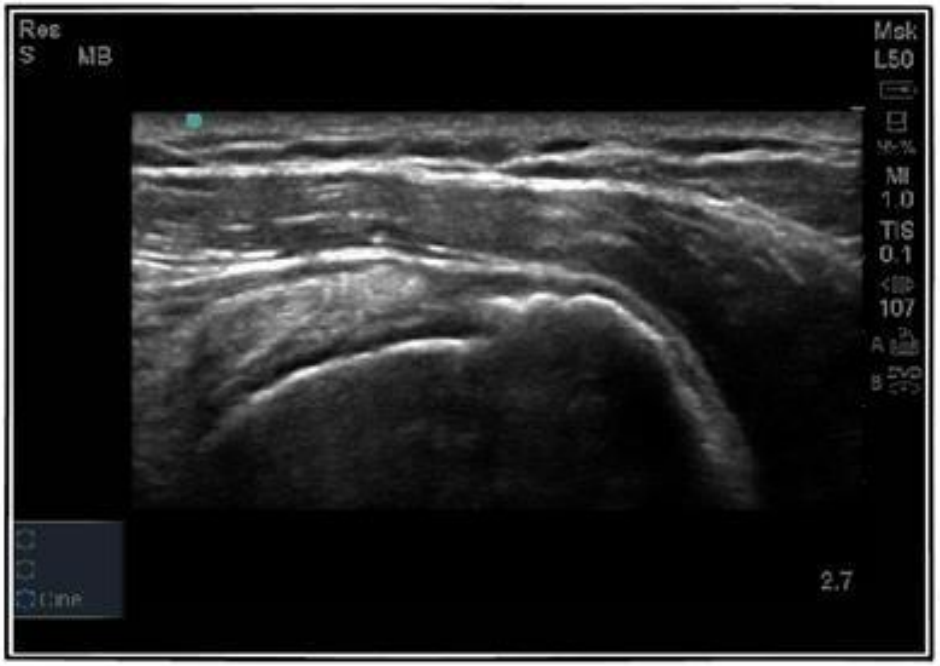
Using your mouse, place the cursor within the region of vessel demonstrating the peak velocity readings; then left-click the mouse to indicate your selection.



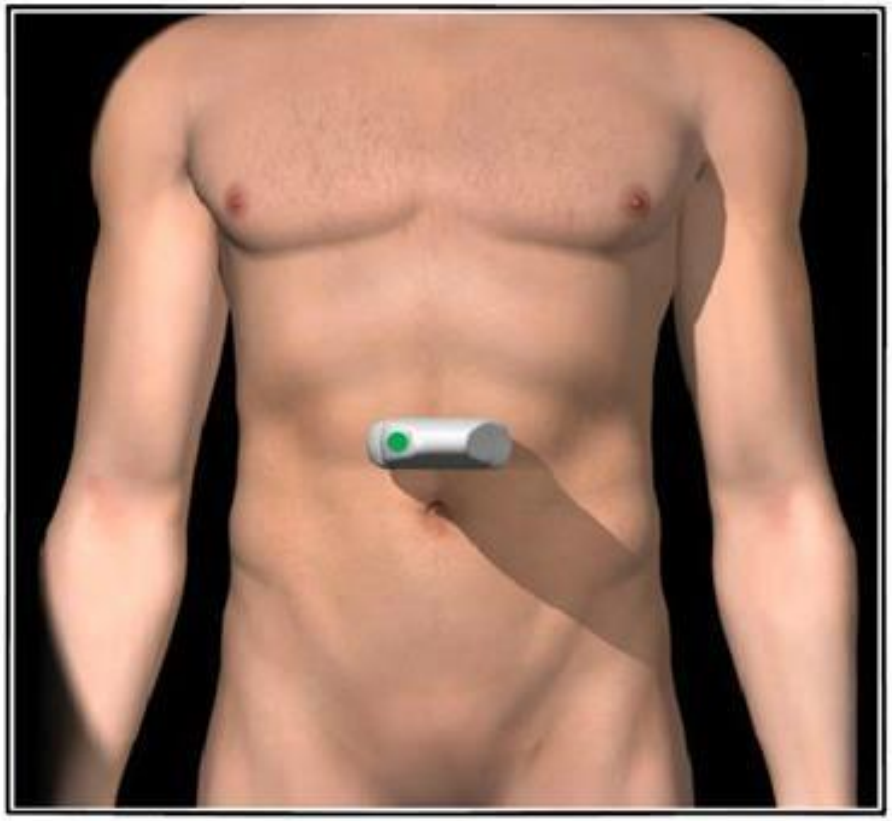
Hot Spot - Single Response, Single Correct

AIT - Avatar

Using your mouse and the control buttons provided, position the patient and transducer appropriately to acquire the image displayed.

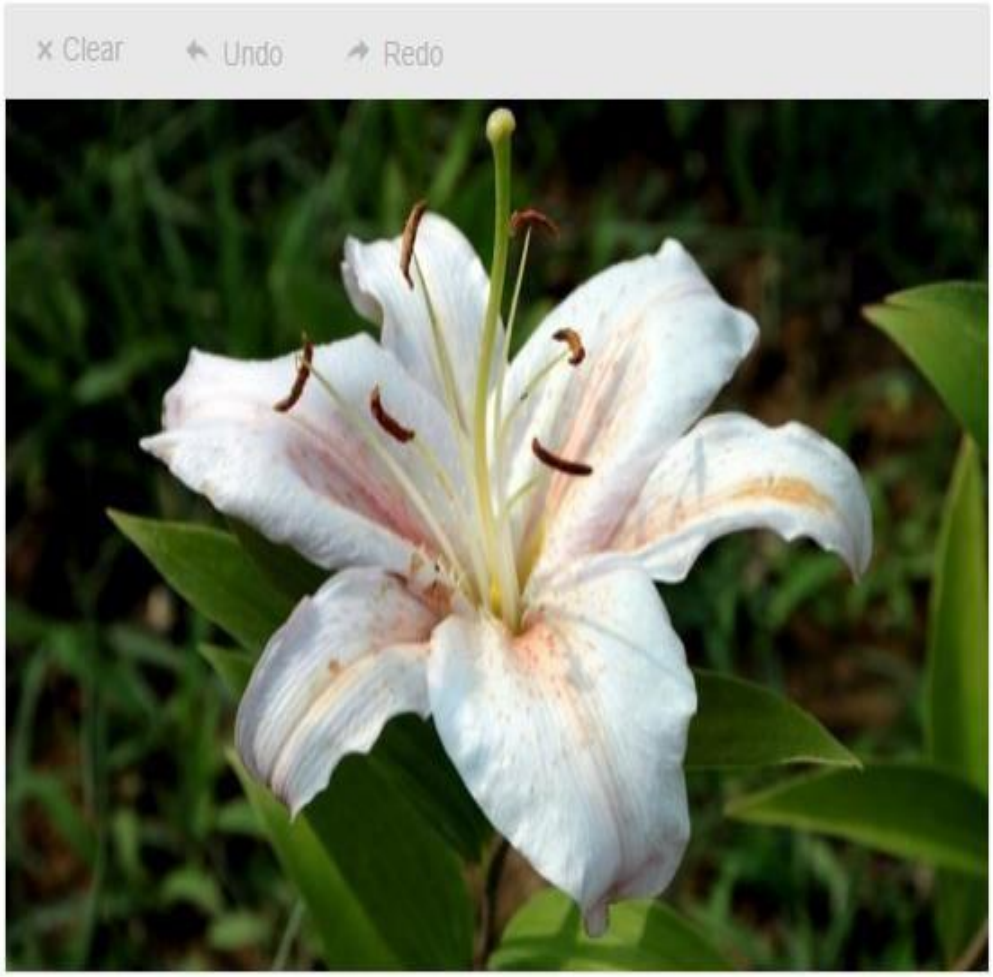


Patient | Transducer



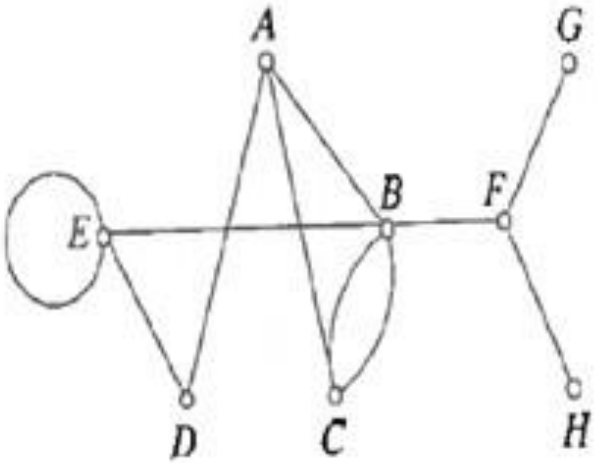
Hot Spot – Single Response, Multiple Correct

Circle one of the flower's anthers in the picture.



Hot Spot – Multiple Response

Click on all vertices that have a degree of 3



Note: There is a designated region around each correct answer within which the test taker can click and get the question correct.

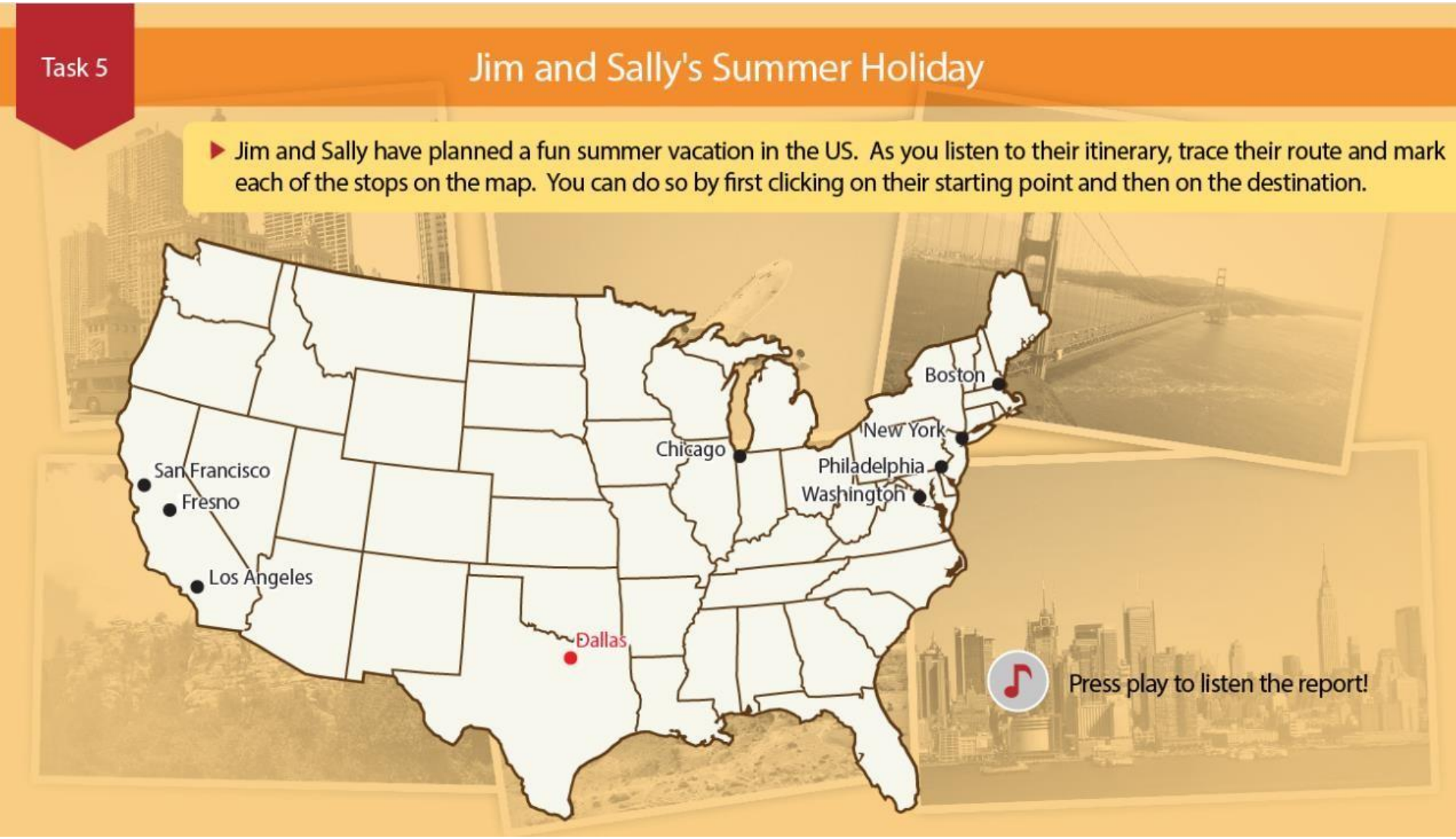
Hot Spot – Multiple Response

Select all the **relevant sections** in the text.

Which sentence or sentences imply that the cheetahs run fast?

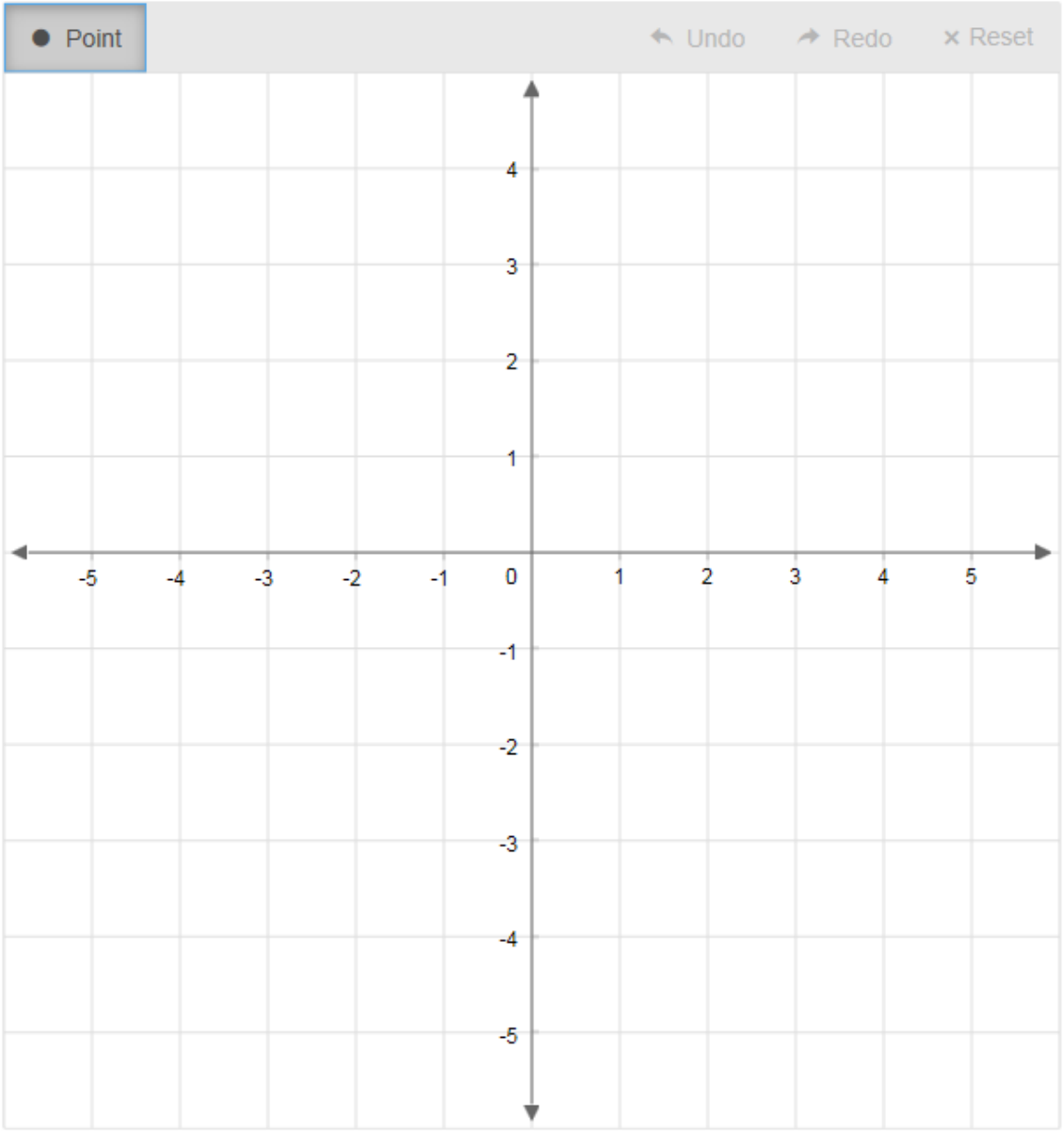
Most cheetahs live in the wilds of Africa. There are also some in Iran and northwestern Afghanistan. The cheetah's head is smaller than the leopard's, and its body is longer. This cat is built for speed. Its legs are much longer than the leopard', allowing it to run at speeds of up to 70 miles per hour! This incredible ability helps the cheetahs catch their dinner, which is usually an unfortunate antelope. A cheetah's spots are simply black spots, not rosettes or circles.

Hot Spot with Audio Prompt



Hot Spot – Plotting Points

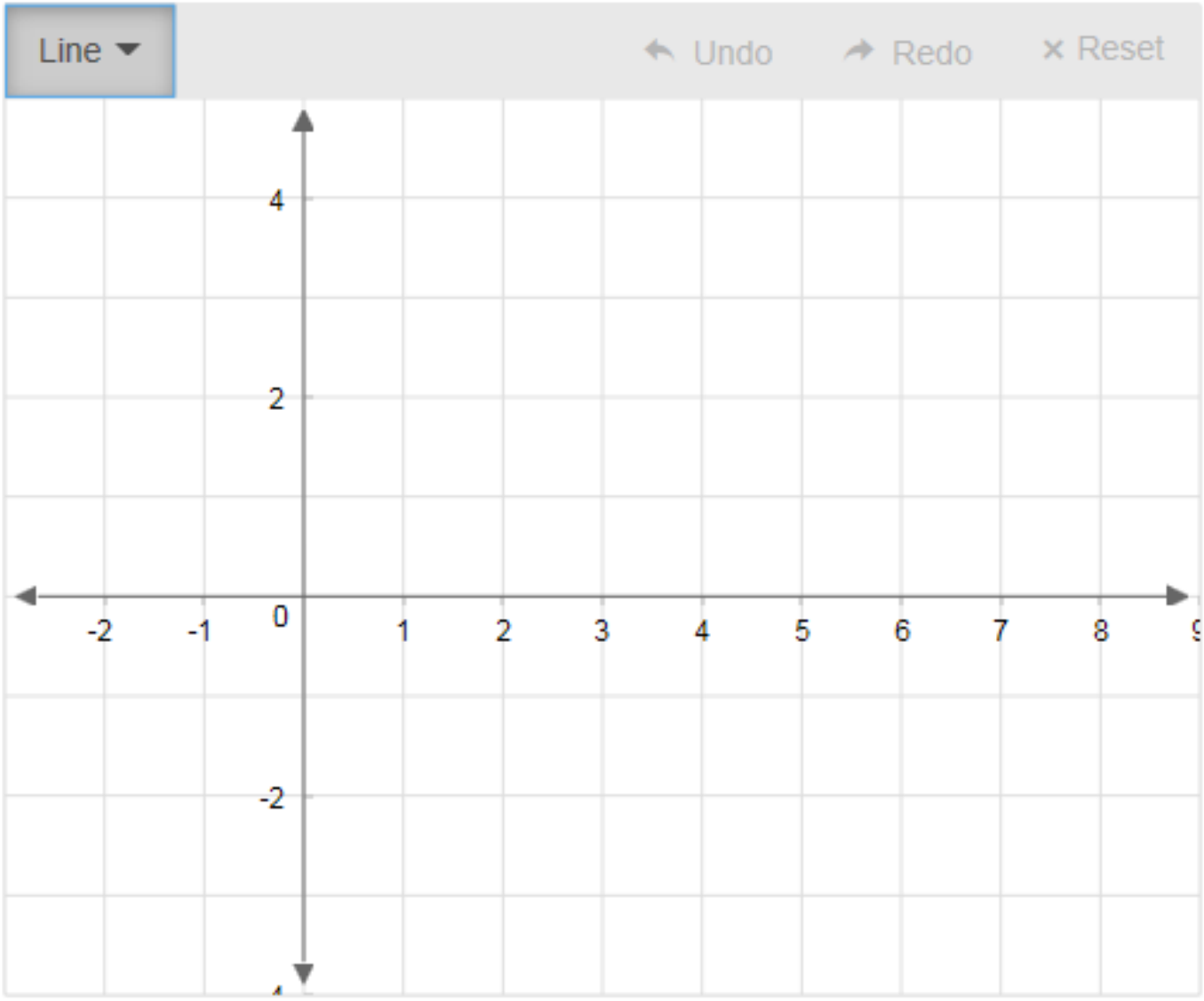
Plot points at $(5,2)$, $(3,0)$, $(2,4)$ and $(-1,-5)$.



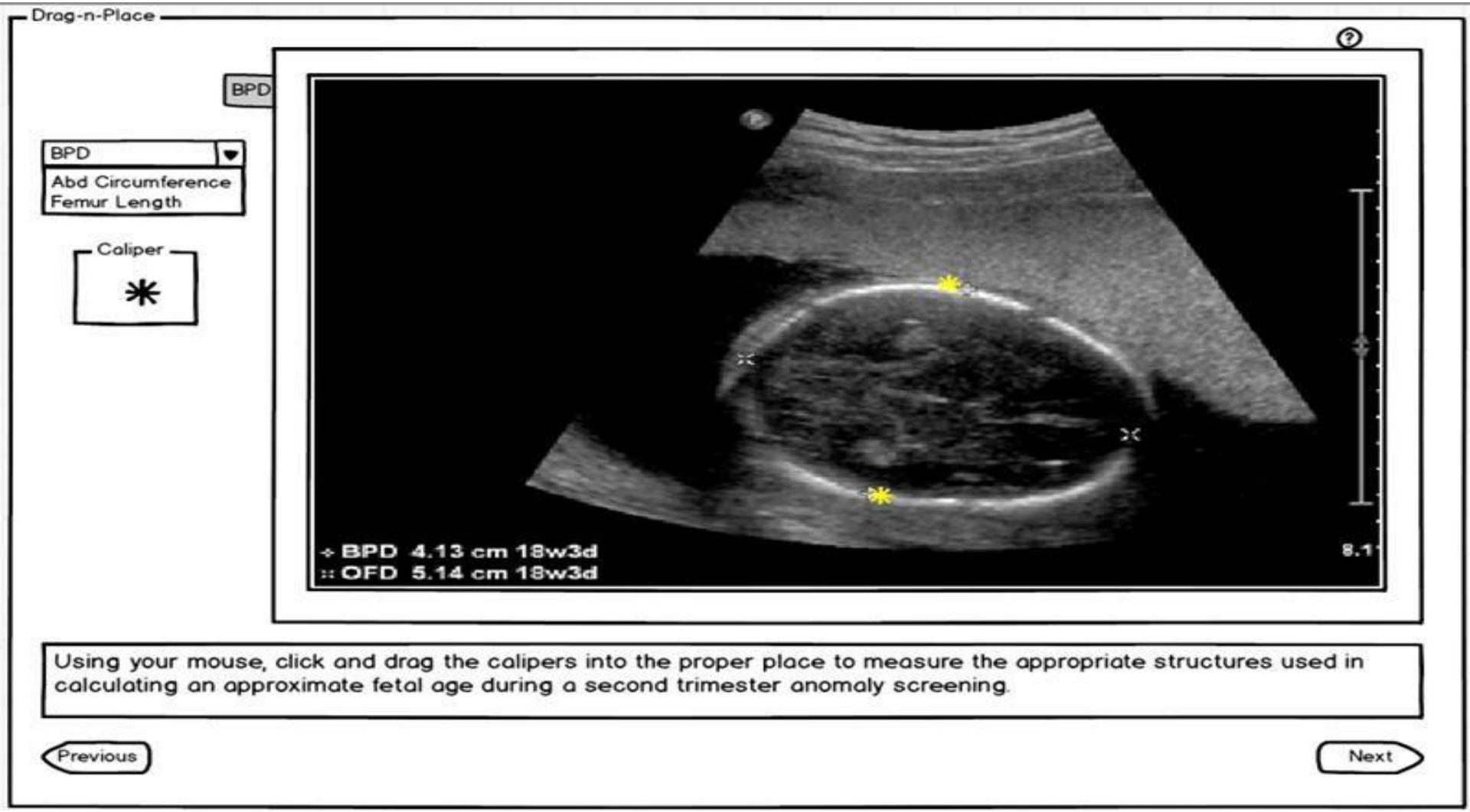
Hot Spot – Plotting Rays

Graph a Ray originating at $(4, 0)$ in the direction towards $(7, 2)$

Hint You'll need to use the **Ray** tool



Hot Spot (presented as drag and drop)



Hot Spot (presented as drag and drop)

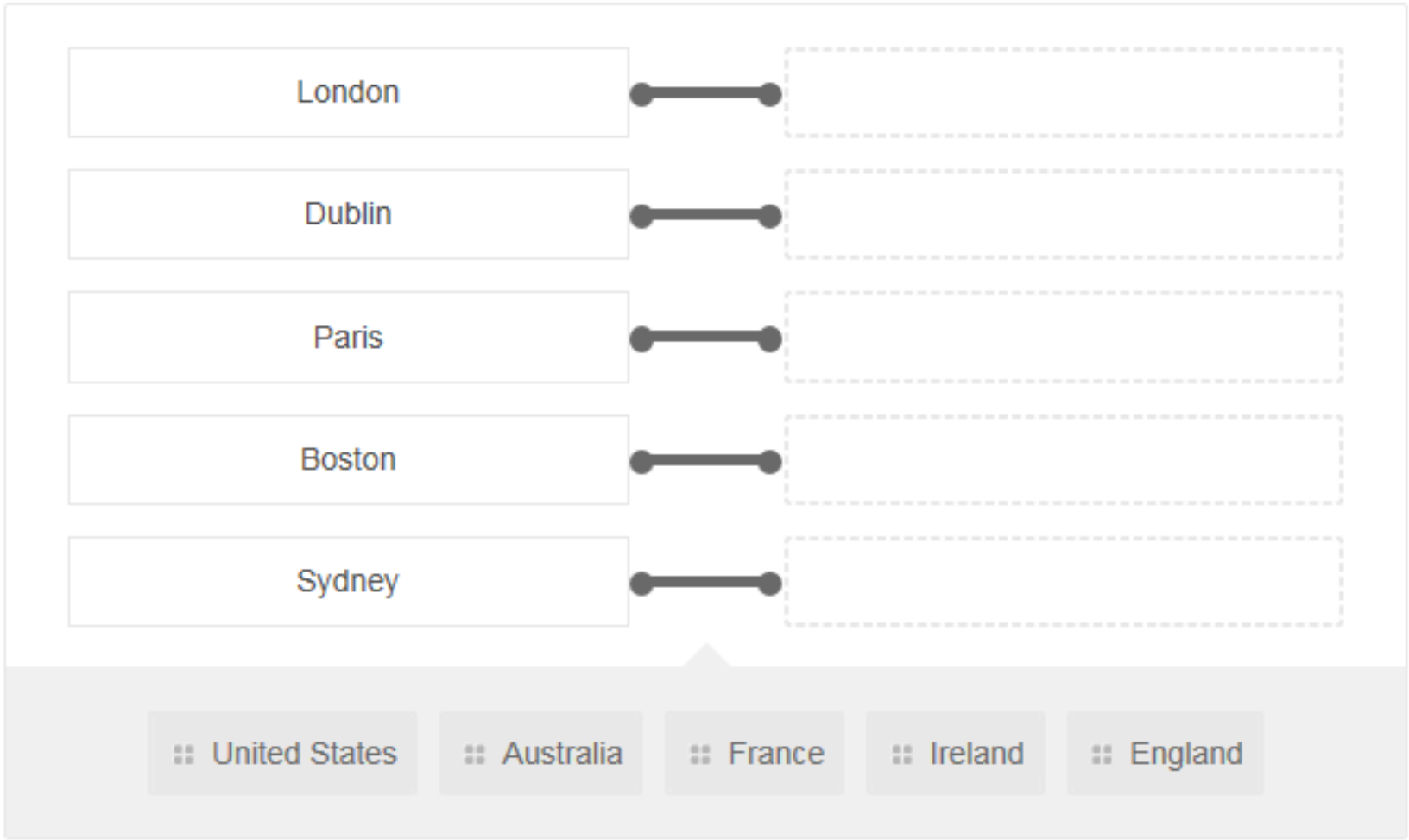
An adolescent transported to the emergency department following a motor vehicle collision complains of shortness of breath.
Utilize a systematic technique to review his chest radiograph, then use your mouse to drag-and-drop the red pin on the MOST significant abnormality.

Hot Spot (presented as drag and drop)

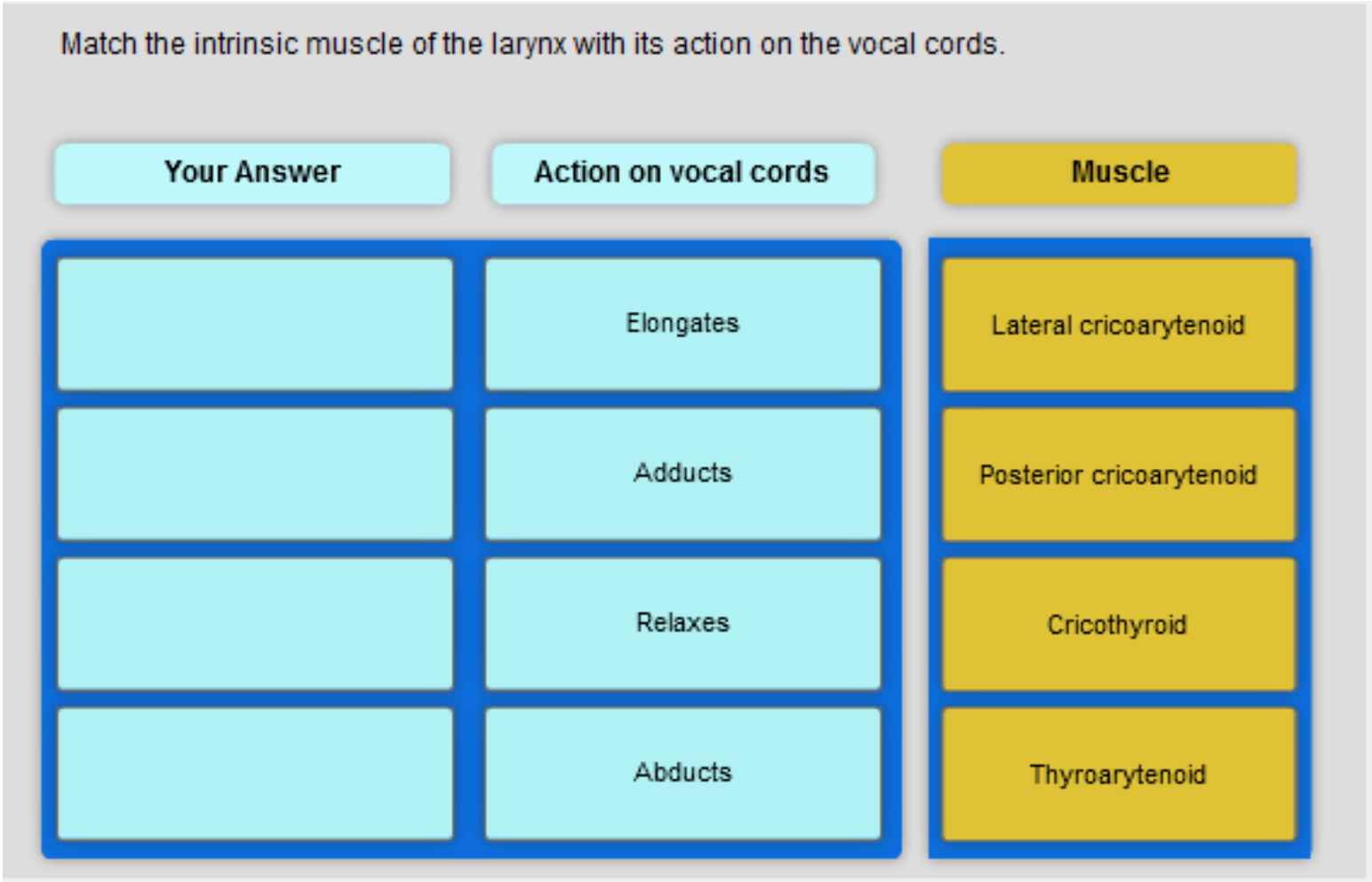
Drag & Drop

Matching

Match each city to its parent nation.



Matching


Match the intrinsic muscle of the larynx with its action on the vocal cords.
Your Answer
Action on vocal cords
Muscle
Elongates
Adducts
Relaxes
Abducts
Lateral cricoarytenoid
Posterior cricoarytenoid
Cricothyroid
Thyroarytenoid

Matching - More Options Than Responses

Task C

Animal symbols in Renaissance painting

Match each animal with its symbolic meaning.

loyalty

strengh

purity

shrewdness

ERMINE

RABBIT

DOG

virginity

laziness

bravery

wantonness

Leonardo da Vinci: Lady with an Ermine

Tiziano Vecellio: The Madonna of the Rabbit

Jan van Eyck: The Arnolfini Portrait

Matching - More Options Than Responses

Match the CSS terms to the corresponding examples. (To answer, drag the appropriate term from the column on the left to its example on the right. Each term may be used once, more than once, or not at all. Each correct match is worth one point.)

Matching -
Options Used Multiple Times or Not at All

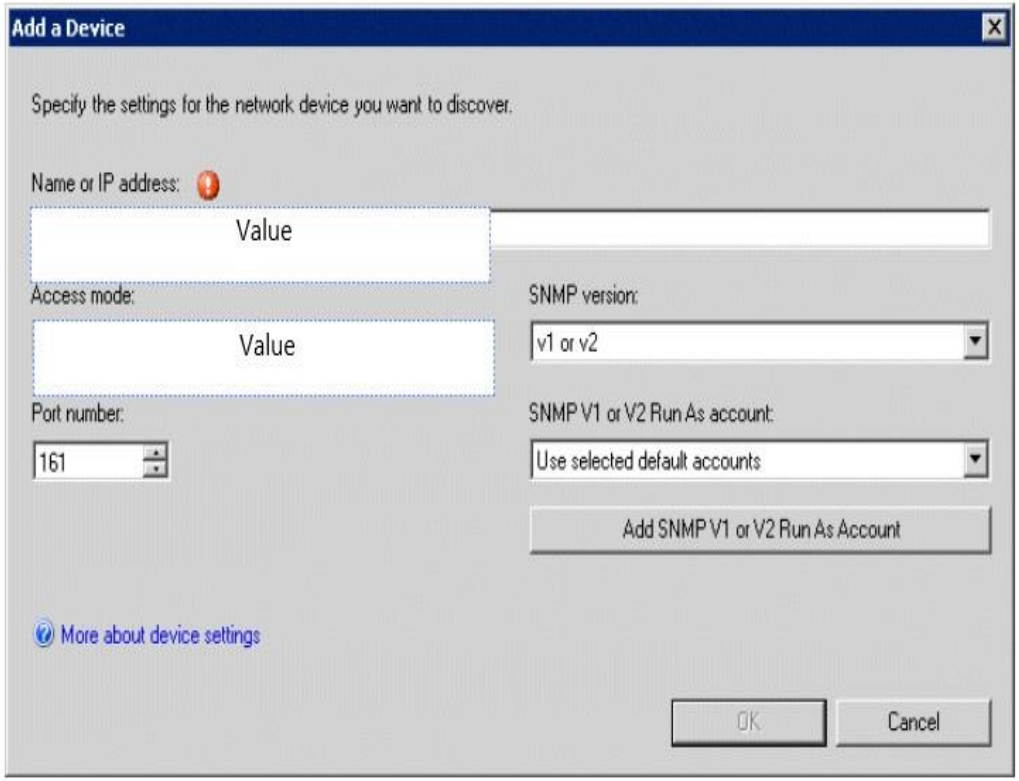
This is the beginning of a question to test the accessibility of a drag and drop item type. Users should be able to begin typing source text into the target area.

To answer, drag the appropriate values to the correct locations in the answer area. (Each value may be used once, more than once, or not at all.) Or, you can type the source text in the target areas.

Values

- ICMP
- SNMP
- ICMP and SNMP
- 10.10.10.12
- 10.10.10.0/24
- Fec0:2308::12

Answer Area



Matching
Options Used Multiple Times

Initial Presentation

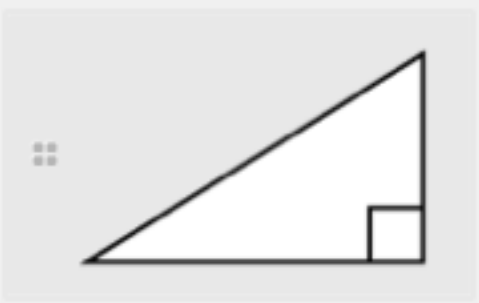
Various geologic rock types are listed at the left. Drag rock types to the left, matching each phrase in the Answer List at the right with the corresponding rock type.

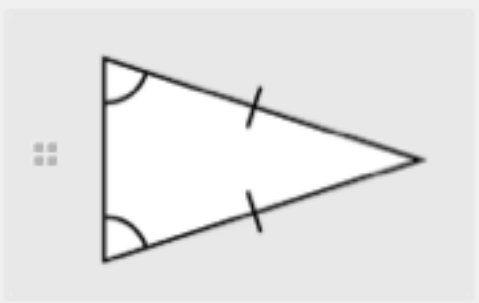
List Items	Answer List
sedimentary	Examples include granite, obsidian, and quartz.
igneous	Formed by deposition, the compaction, and finally cementation.
metamorphic	Examples include chalk, coal, and shale.
	Formed from the other rock types.
	The two main categories are intrusive and extrusive.
	Formed by the cooling of molten magma.
	Fossils are mostly found in this rock type.
	Examples include schist, magnetite, and graphite.

Matching – Graphics as Options

Drag each triangle to the correct category.

Isoceles	Scalene	Equilateral





Matching – Graphics as Options

Matching - Sentence Completion

Task G

Renaissance Statuary

Complete the sentences below by adding the names of artists and cities in the right place.

Venice

Michelangelo

Florence

Donatello

Padue

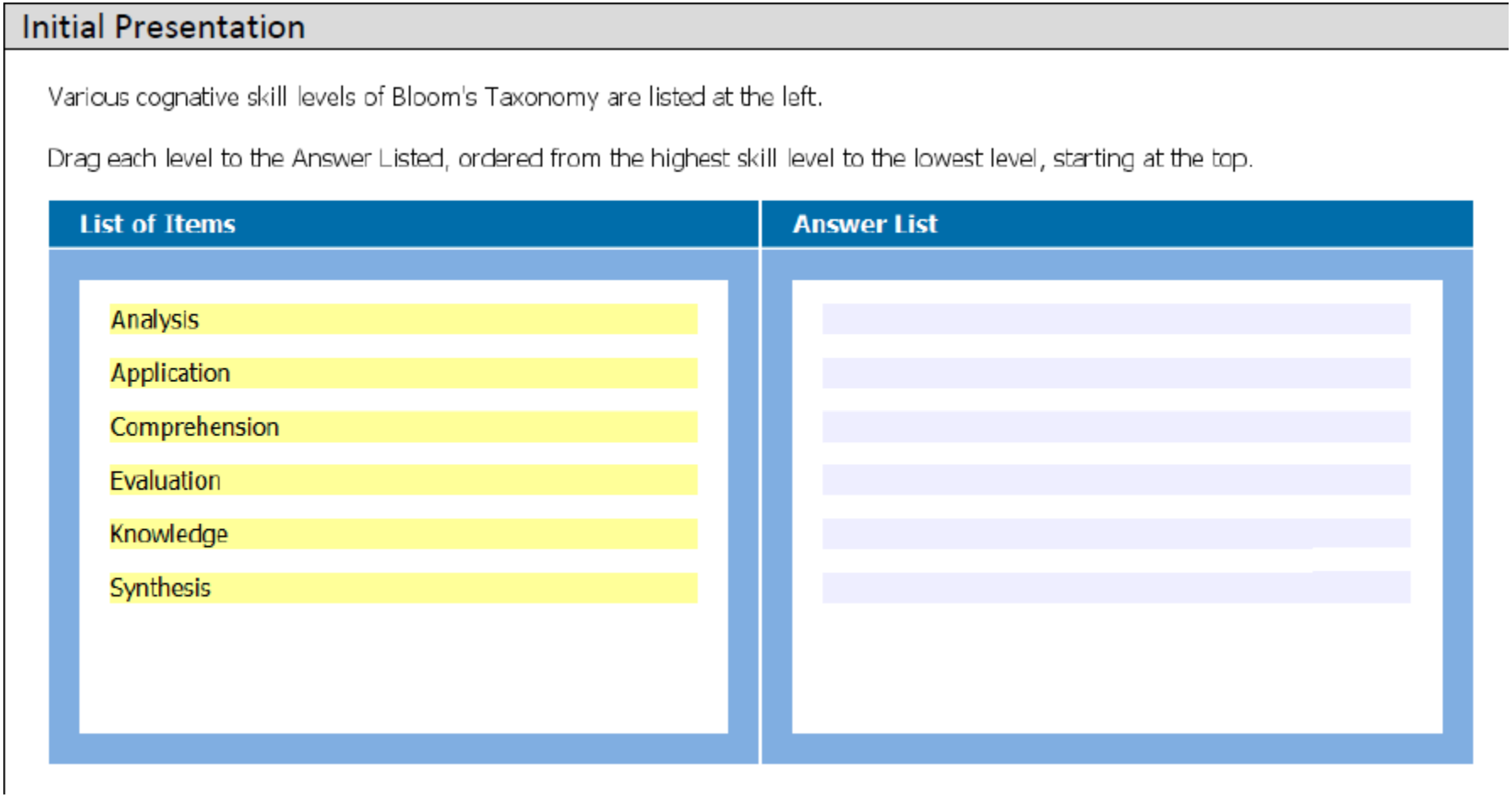
Verrocchio

_____ 's statue of David can be found in _____

_____ 's equestrian statue of Gattamelata can be found in _____

_____ 's statue of Captain General Colleoni can be found in _____

Ordering/Ranking



Ordering/Ranking

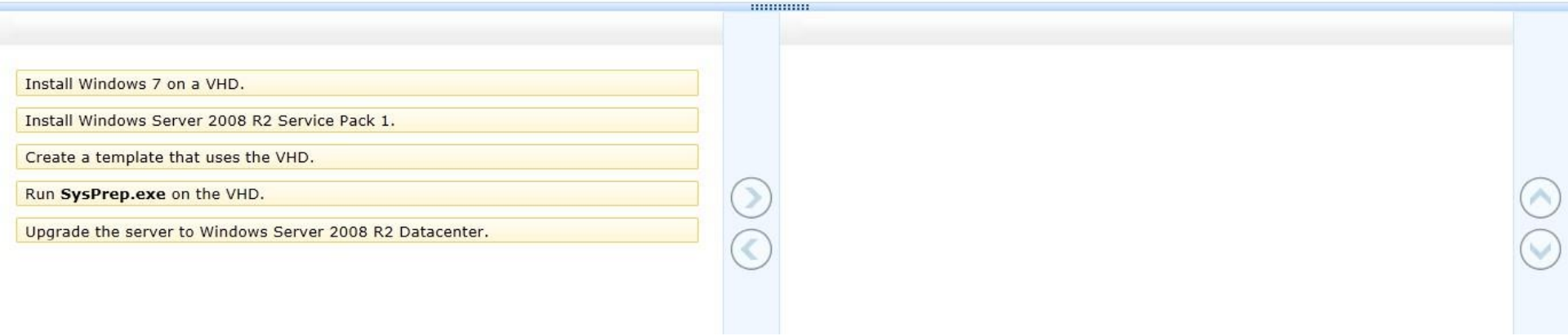
A company has a server that runs Microsoft System Center Virtual Machine Manager (VMM) 2008 R2 with Service Pack (SP) 1 and Windows Server 2008 R2 Enterprise with Hyper-V.

The company is preparing to deploy virtual machines (VMs) from templates and has the following requirements:

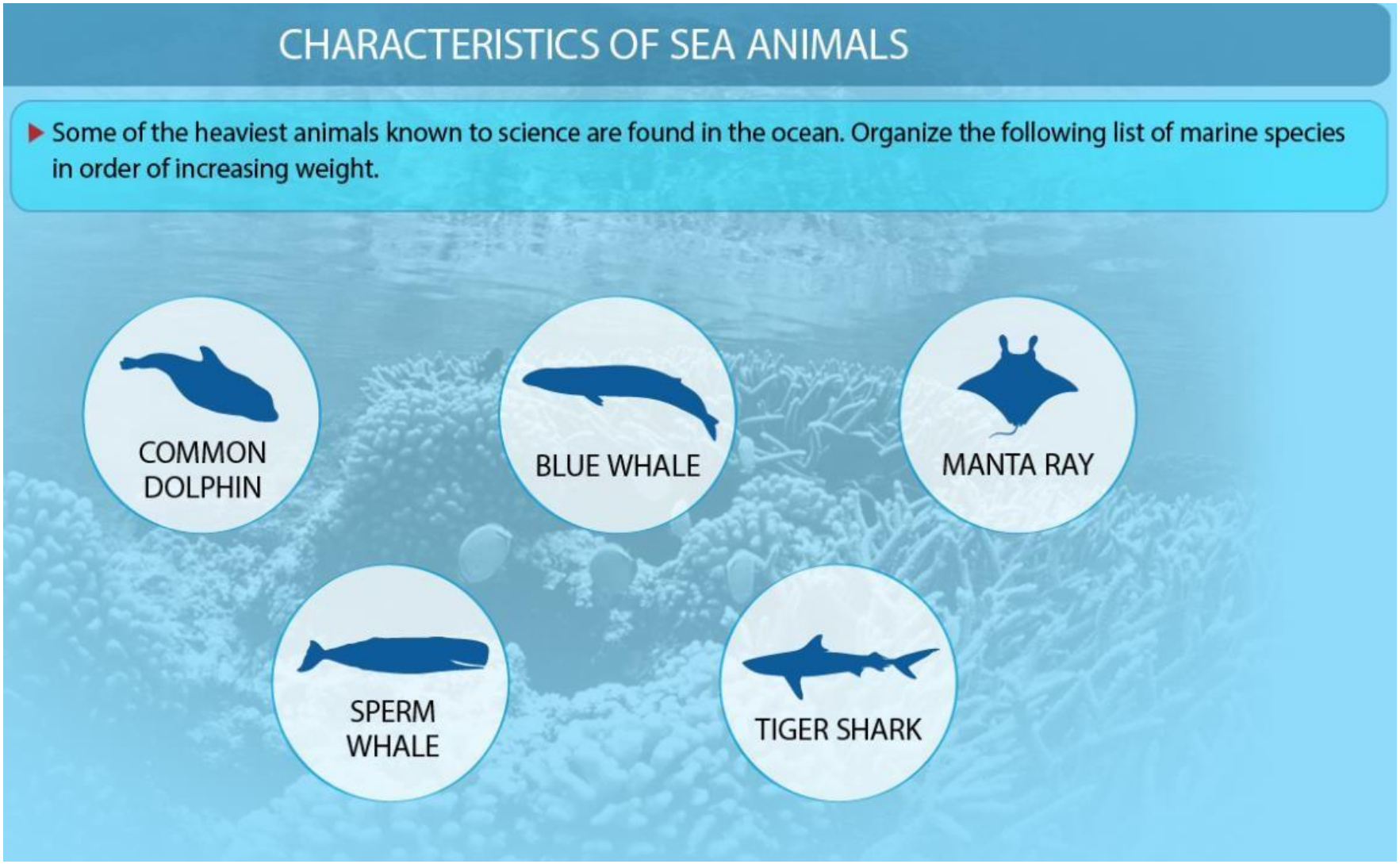
- The templates must be created from virtual hard disks (VHDs).
- The templates must include Windows 7.
- An out of the box experience (OOBE) must be provided for all guest operating systems that are deployed from the templates.

You need to create a template that meets the company requirements.

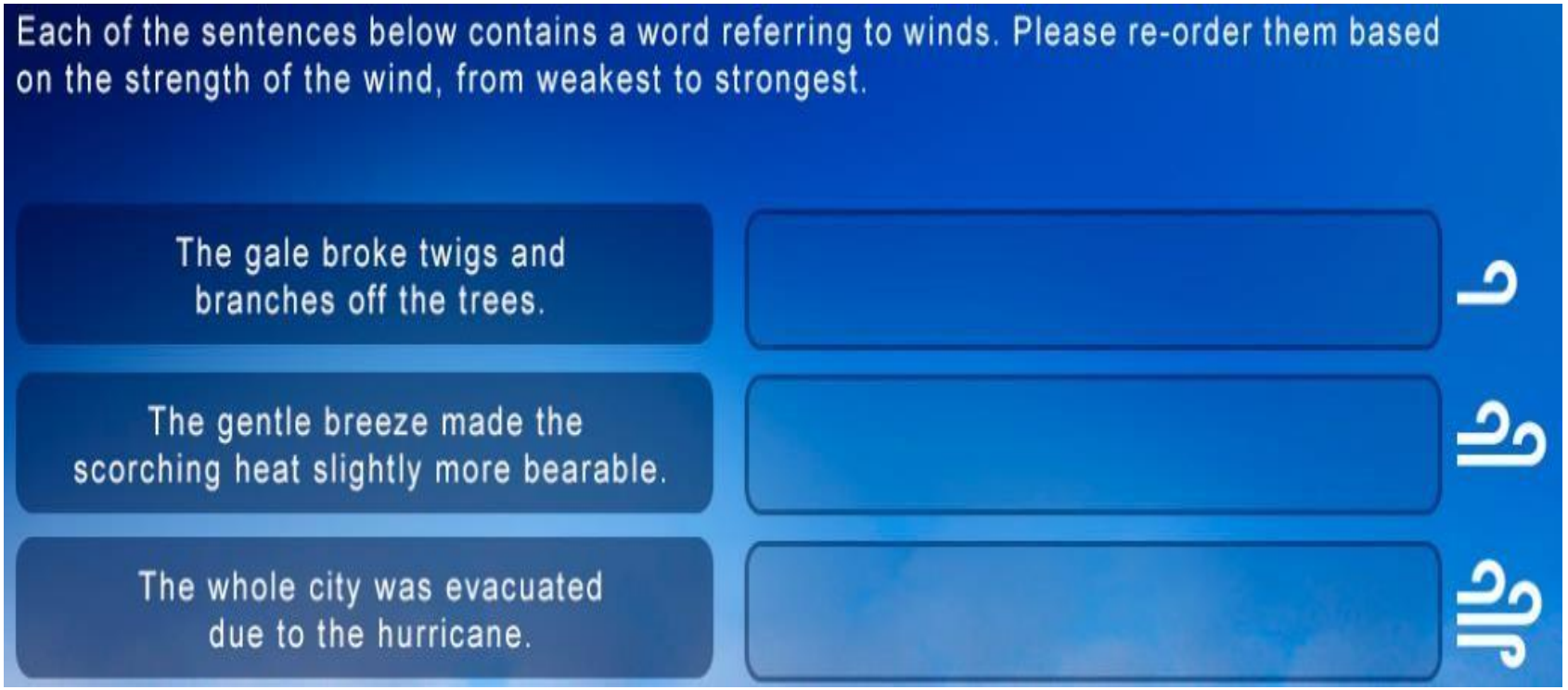
Which three actions should you perform in sequence? (To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.)



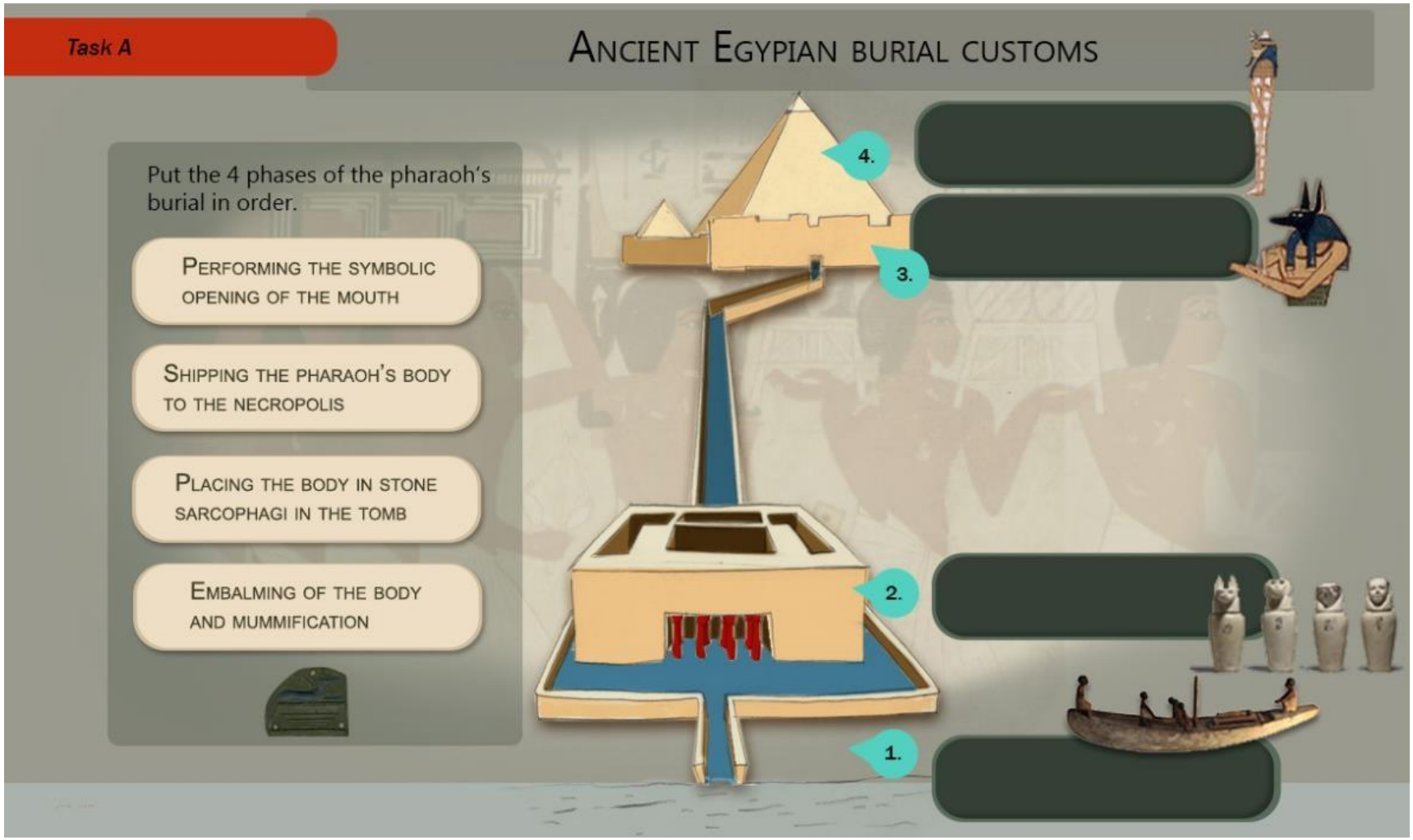
Ordering/Ranking



Ordering/Ranking



Ordering/Ranking



Ordering/Ranking

You are creating a Win8 application that enables users to manage files.

You need to enable users to select a single file. (Develop the solution by arranging the code. You will need all of the code blocks.)

```
OutputTextBlock.Text = "Operation cancelled.";
    }
```

```
if (rootPage.EnsureUnsnapped())
{
```

```
FileOpenPicker openPicker = new FileOpenPicker();
    openPicker.ViewMode = PickerViewMode.Thumbnail;
    openPicker.SuggestedStartLocation =
PickerLocationId.PicturesLibrary;
    openPicker.FileTypeFilter.Add(".jpg");
    openPicker.FileTypeFilter.Add(".jpeg");
    openPicker.FileTypeFilter.Add(".png");
```

```
StorageFile file = await openPicker.PickSingleFileAsync();
    if (file != null)
```

```
{
```

Answer Area

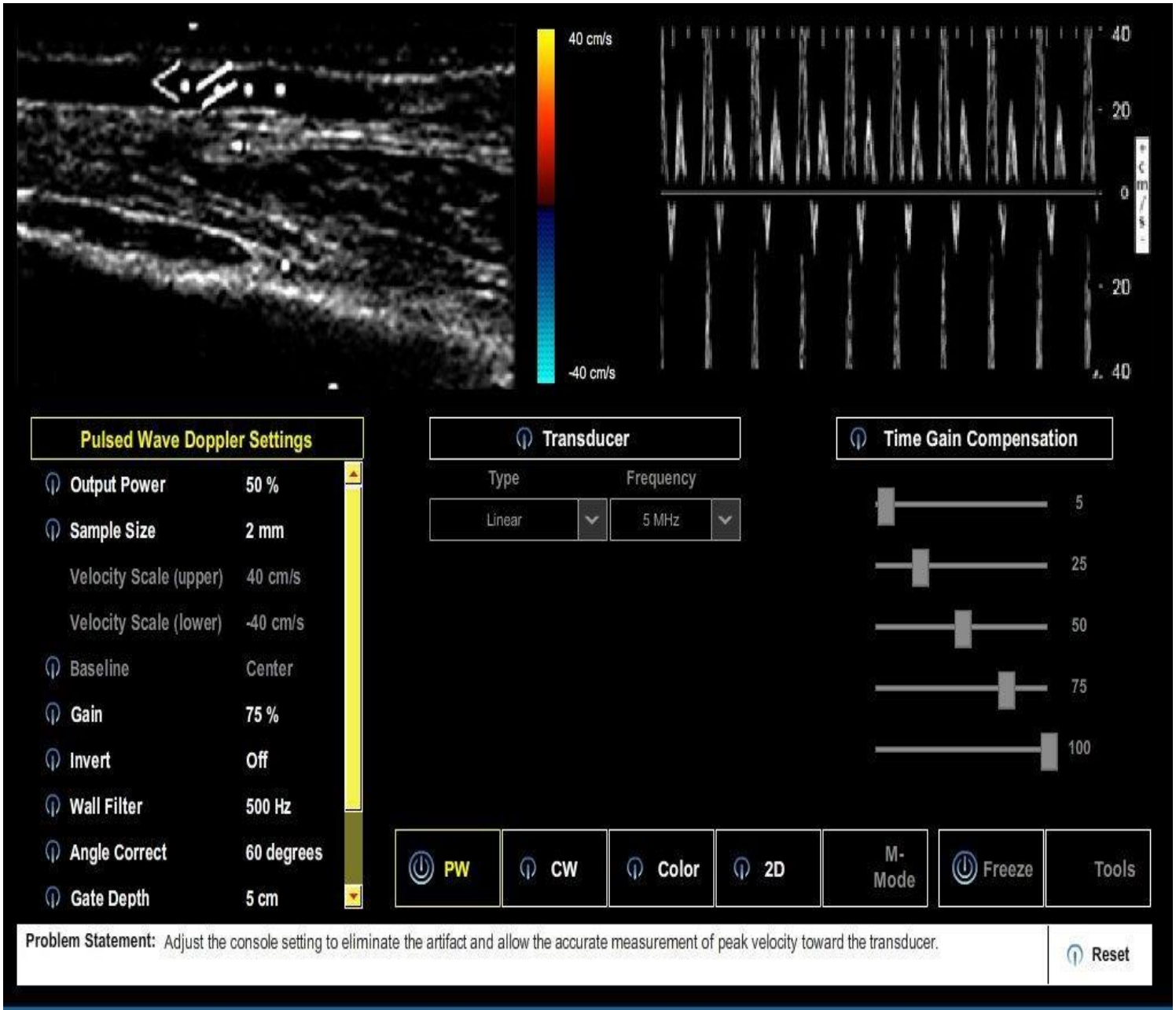
Ordering/Ranking

Rearrange the sentences into the correct order.

≡ On the contrary, for a small street in a quiet neighbourhood, it was remarkably animated. ≡ There was a group of shabbily dressed men smoking and laughing in a corner, a scissors-grinder with his wheel, two guardsmen who were flirting with a nurse-girl, and several well-dressed young men who were lounging up and down with cigars in their mouths. ≡ It was a quarter past six when we left Baker Street, and it still wanted ten minutes to the hour when we found ourselves in Serpentine Avenue. ≡ The house was just such as I had pictured it from Sherlock Holmes' succinct description, but the locality appeared to be less private than I expected. ≡ It was already dusk, and the lamps were just being lighted as we paced up and down in front of Briony Lodge, waiting for the coming of its occupant.

Simulation

Semi-Interactive Console



About Semi-Interactive Console Items

- The problem statement is at the bottom of the screen.
- Test takers can click on different areas of the console and make adjustments.
- Specific settings (possibly more than one set is acceptable) are required to get the item correct.
- The image is frozen and does not change with adjustments.

Interactive Spreadsheet

Cut | Copy | Paste

Sales and production costs for a company's product are provided below.

Calculate the percent change in gross profit per unit if the price of shipping decreases by 50% (round to the nearest % and state as an absolute value/positive number). Also, indicate the direction of the movement (increase or decrease).

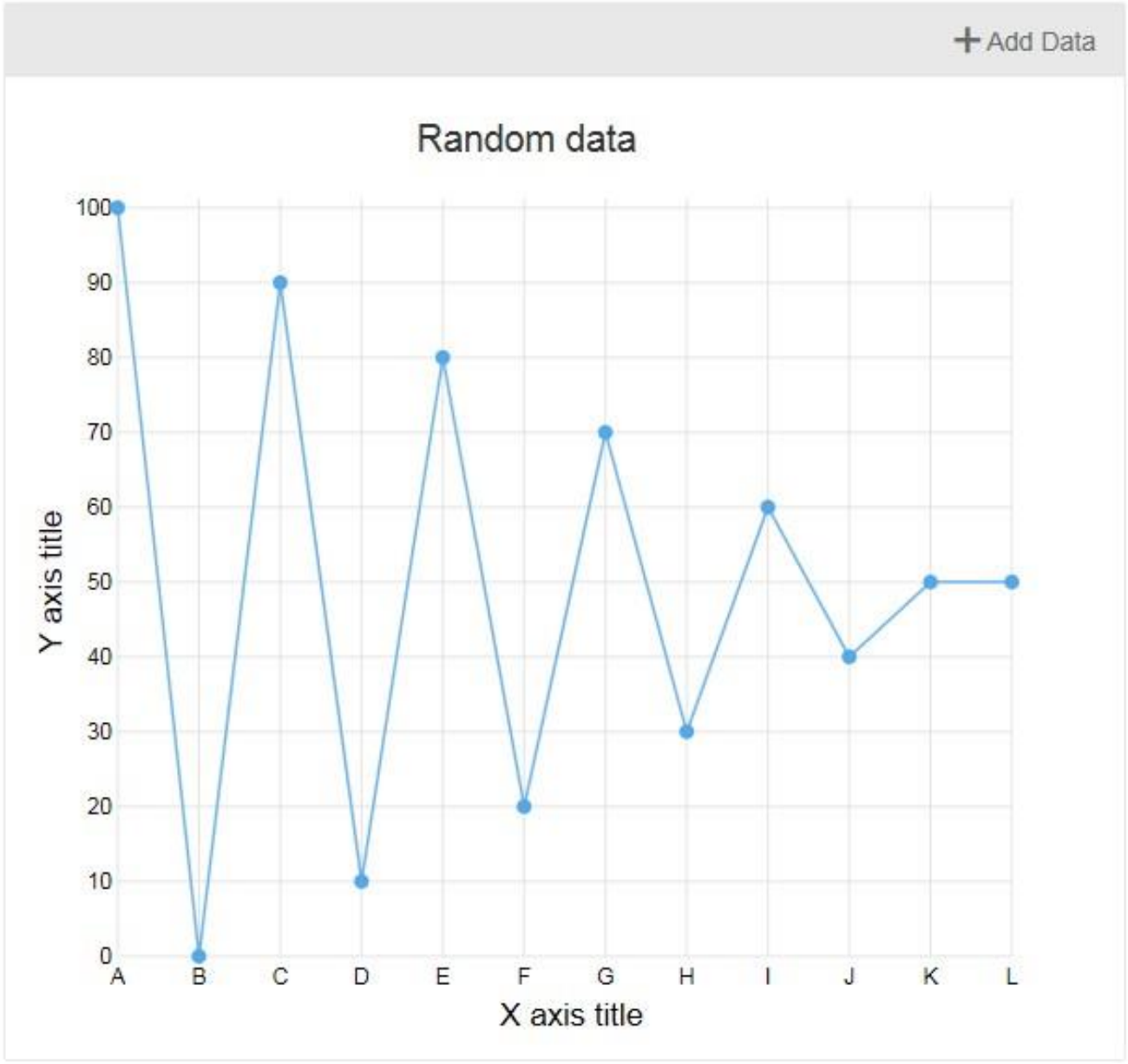
C2 ✖ ✔ *fx*

	A	B	C	D	E	F
1						
2		**Change (%)**				
3						
4		**Direction**				
5			(double-click)			
6						
7		Sales (units)	1,000			
8		Sales price (per unit)	$10			
9		Production costs				
10		Labor (per unit)	$5			
11		Materials (per unit)	$3			
12		Shipping (per unit)	$1			
13						
14						

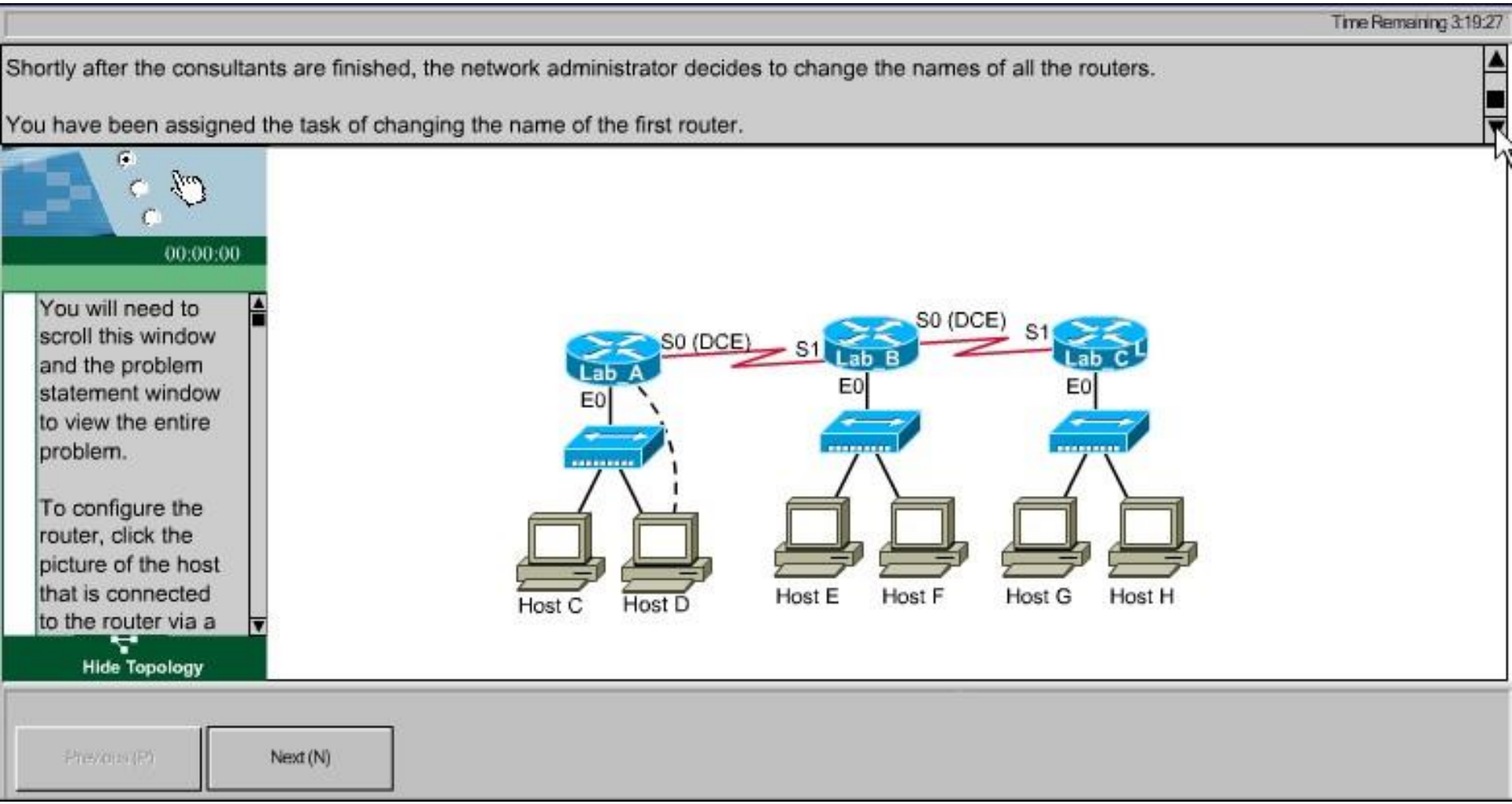
Test-takers are given a scenario and can use a spreadsheet formulas to calculate the correct answer(s). The spreadsheet resembles Excel, but has more limited functionality.

Interactive – Line Chart

Hint Resize L to 40 and add a new point (M) and set its value to 60.

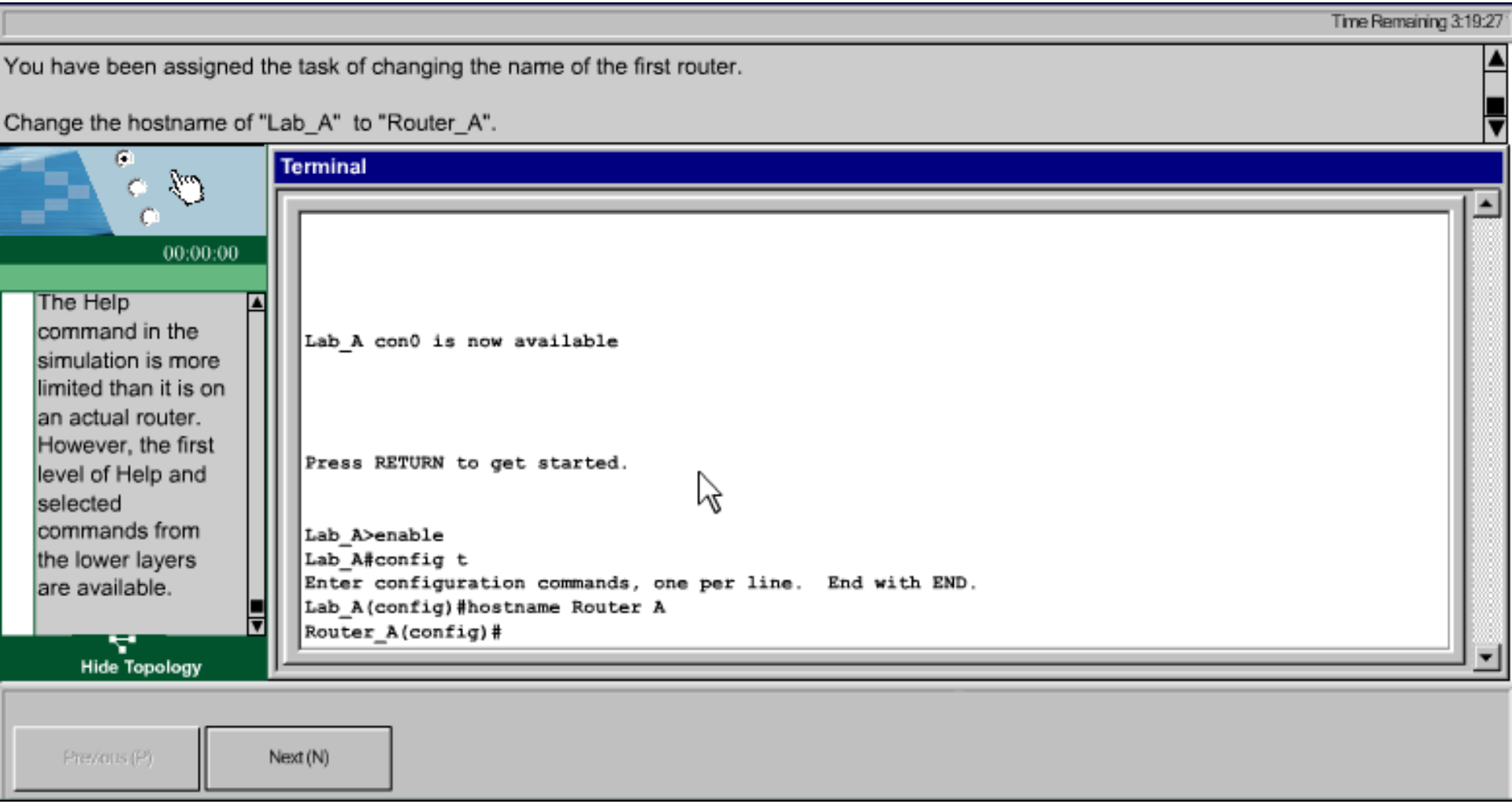


Code Simulation (1 of 2)



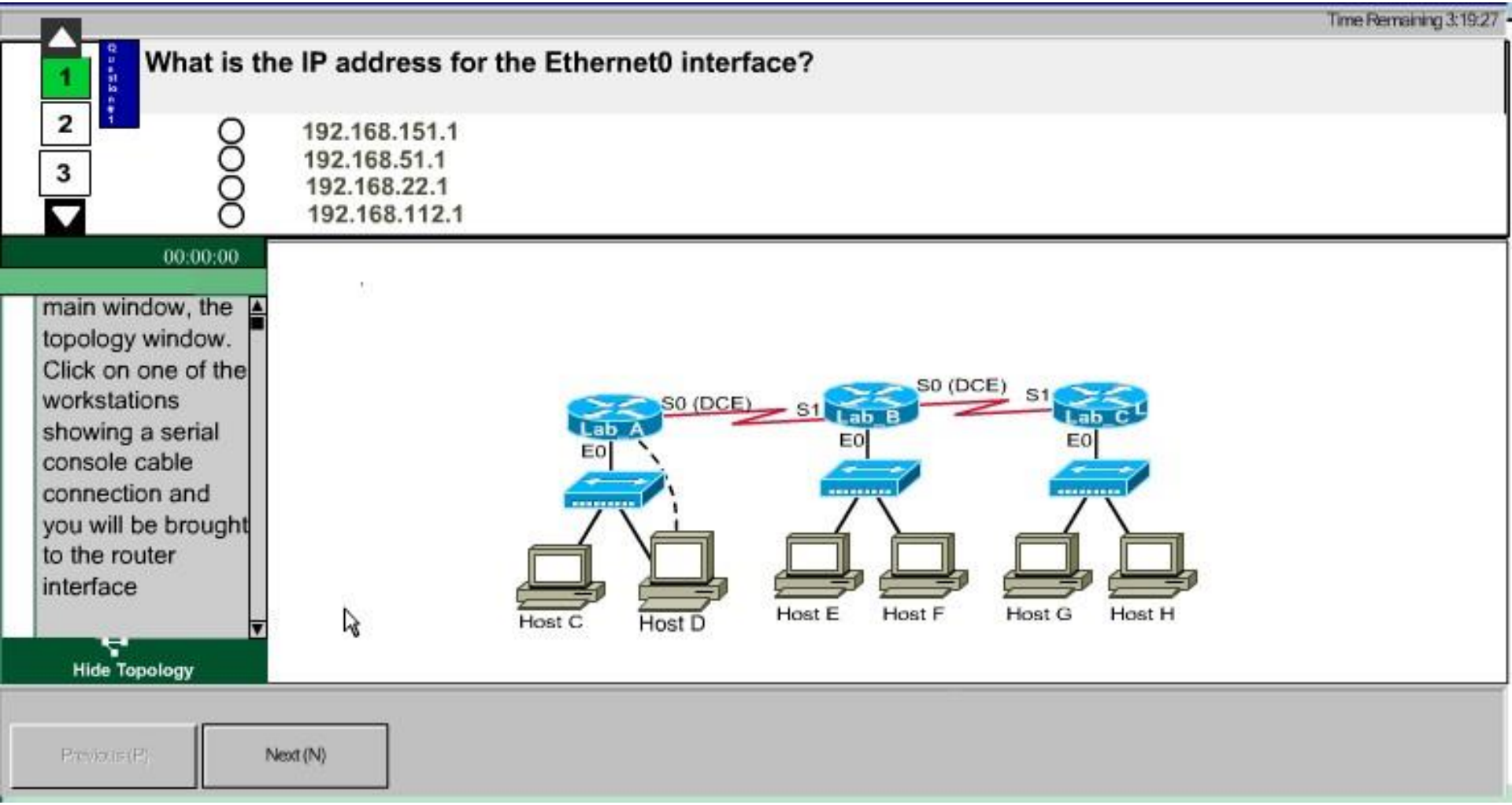
Test taker reads scenario provided at top of the screen and instructions on left (both require scrolling). Then, the test taker clicks on graphic to open terminal (see next slide).

Code Simulation (2 of 2)



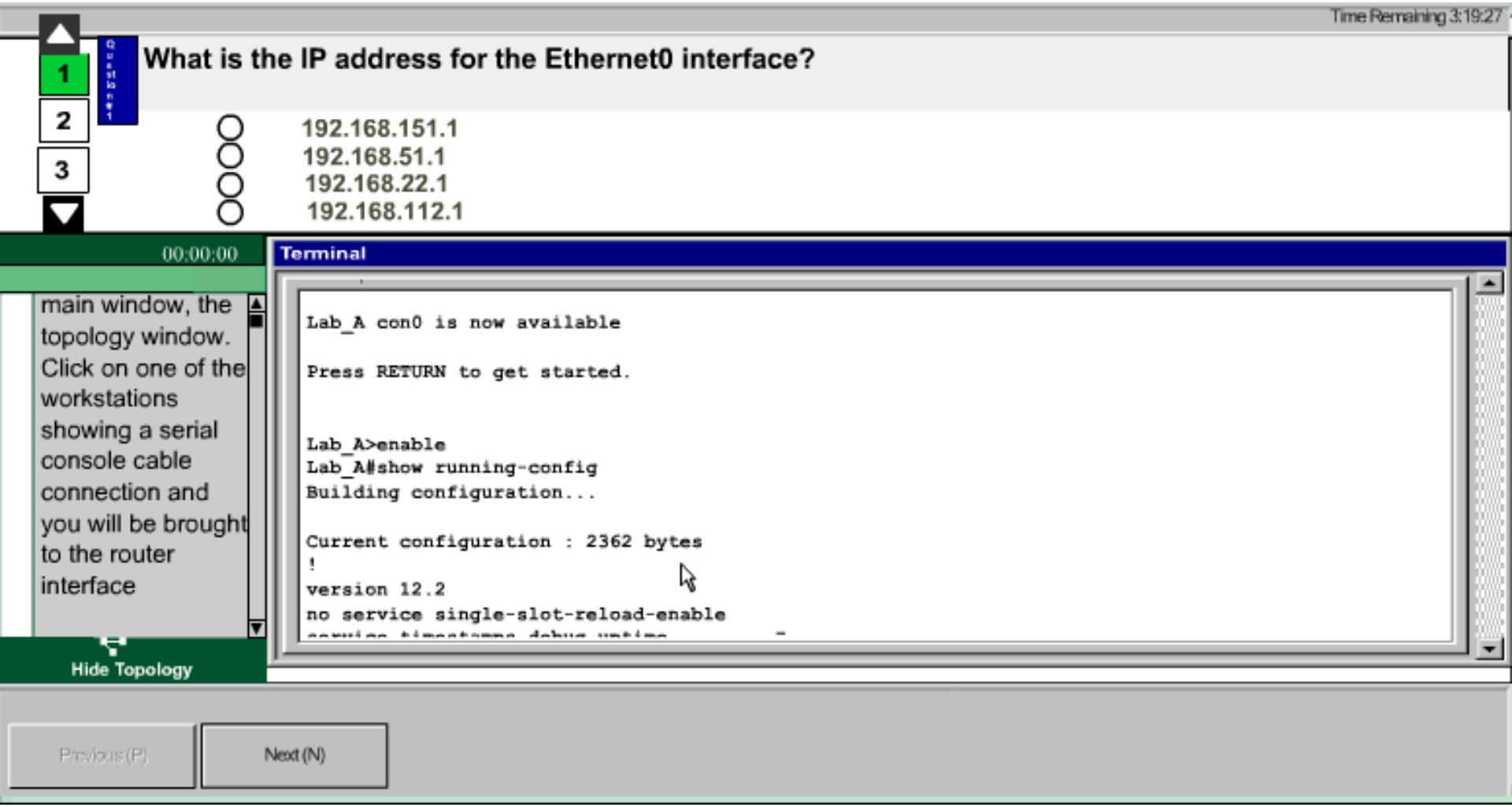
Test taker writes code in the terminal to complete prompt in stem.

Mini Simulation with MCQs (1 of 2)



Test taker reads the instructions on the left side of screen (requires scrolling), then clicks on one of the workstations.

Mini Simulation with MCQs (2 of 2)



Test taker types code in the Terminal to find the answers to the three MCQs at the top of the screen.

Simulation with MQCs

Test taker reviews scenario and multiple images and videos to answer the MCQs with drop-down options.