# 00 Intro Page

|  |  |
| --- | --- |
| Page Title in Header | Assessments for a Changing Education Landscape |
| Header Text | 21st century educators and policy makers are no longer just interested in what students know. They want to understand what students can do with what they know. To assess this, many are implementing performance assessments. |
| Body Header | But what is “performance assessment” and how is it different from other types of assessments? |
| Body Copy | The Center for NextGen Learning & Assessment has developed a Framework of Approaches to Performance Assessments to answer that question. This framework provides a detailed description of seven approaches to performance assessment. It also includes innovative examples from a variety of sources to create a foundational understanding among policy makers, educators, parents, and even students about these assessments. |
| Call Out | All students should have access to engaging learning that helps them apply complex 21st century skills and prepare for brighter futures. Performance assessment that is linked to rigorous curriculum and instruction helps drive more comprehensive learning and measure achievement as required by the Common Core State Standards (CCSS). |
| Body Copy | The Center for NextGen Learning & Assessment conducts research on new methods and technologies to support students, teachers, and educational decision makers with more realistic, engaging learning and assessment experiences. Current core topics are to increase capabilities, quality, and efficacy in the areas of performance assessment and automated scoring.  For more information on our team or the research we are conducting, visit our [website](http://researchnetwork.pearson.com/nextgen-learning-and-assessment). |

# 00 Homepage

|  |  |
| --- | --- |
| Page Title in Header | Purpose of the Framework |
| Header Text |  |
| Body Header |  |
| Body Copy | This framework is designed to:   * Reduce ambiguity in the education community around what performance assessments are and how they may be used * Share research-based descriptions that define and differentiate each of the seven approaches * Provide resources and considerations to use when making decisions about performance assessments * Showcase a variety of performance assessment examples, including those developed internationally and for purposes outside of K-12 education |
| Graphic | PerformanceAssessmentApproachs\_2.2.png |
| Brainshark | Yes |

# 01 What is Performance Assessment

|  |  |
| --- | --- |
| Page Title in Header | What is Performance Assessment? |
| Header Text | An assessment activity or set of activities that requires examinees to generate products or performances that provide direct or indirect evidence of the examinee’s knowledge, skills, and abilities (KSAs) in an academic content domain or professional discipline. |
| Body Header | Context for Use |
| Body Copy | Typically, performance assessments emulate a context outside of the assessment in which the KSAs ultimately will be applied; require use of complex knowledge, skills, and/or reasoning; and require application of evaluation criteria to determine levels of quality, correctness, or completeness. |
| Body Header | Purposes |
| Body Copy | Performance assessments support both formative and summative uses and are believed to be useful for guiding instruction and encouraging educators to increase expectations for student learning. Proponents argue that performance assessments constitute more direct evidence of what students know and can do than more constrained assessment approaches. |
| References | American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1999). St*andards for educational and psychological testing.* Washington, DC: American Educational Research Association.  Ferrara, S. (1993). Generalizability theory and scaling: Their roles in writing assessment and implications for performance assessments in other areas. Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta, GA.  Fitzpatrick, R., & Morrison, E. J. (1971). Performance and product evaluation. In R. L. Thorndike (Ed.), *Educational measurement*(2nd ed., pp. 237-270). Washington DC: American Council on Education.  Herman, J. L., Aschbacher, P. R.  & Winters, L. (1992). A practical guide to alternative assessment. Alexandria, Va.: Association for Supervision and Curriculum Development.  Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement:* Issues and Practice, 18(2): 5-17. |
| Graphic | PerformanceAssessmentApproachs\_2.2.png |
| Video | Yes |

# 04 Short Constructed Response & Technology-Enhanced Items

|  |  |
| --- | --- |
| Page Title in Header | Short Constructed-Response & Technology-Enhanced Items |
| Header Text |  |
| Body Header | Short Constructed-Response |
| Body Copy | Short constructed-response (SCR) items, whether paper-based or online-administered, require examinees to develop a partial or full response to a stimulus or prompt as opposed to selecting from a limited set of pre-specified options.  In general, features of SCR items include:   * Elicit textual or graphical responses that are discrete and brief * Often allow multiple appropriate answers and/or partial credit scoring * May better support diagnosis and remediation of certain specific skill deficiencies than selected-response items * May be human or machine scorable |
| Body Header | Technology-Enhanced Items |
| Body Copy | Technology-enhanced items (TEIs) feature technology enhancements— such as the addition of interactive functionality, sound, graphics, animation, or video—intended to substantially improve some aspect of measurement.  In general, features of TEIs include:   * Elicit responses that are discrete and brief * Are typically machine scorable * Expand the range of content, knowledge, or skills that are assessable in large-scale contexts compared to traditional paper-and-pencil formats |
| Brainshark | Yes |

## 04a More Info

|  |  |
| --- | --- |
| Page Title | What Are Short Constructed-Response and Technology-Enhanced Items? |
| Definition | **Short constructed-response (SCR) items,** whether paper-based or online-administered, require examinees to develop a partial or full response (textual or graphical) to a stimulus or item stem as opposed to selecting from a limited set of pre-specified options. These items elicit discrete and brief responses, in contrast to more extended constructed responses elicited with other assessment approaches, such as essay prompts.  SCRs often support multiple appropriate answers, both with and without partial credit scoring. As a result of these features, SCR items enable more direct measurement of certain types of knowledge and skills than selected-response (SR) items. Because these items are less constrained than SR items, they often appear more similar to the types of open-ended problems typically encountered in academic and work settings (i.e., they offer more realistic or authentic problem representations). SCRs may provide more detailed information about specific error tendencies than SR items; as such, they may better support diagnosis and remediation of specific skill deficiencies. Finally, certain types of SCRs (e.g., computer-administered items that call for a graphical response) are machine scorable, which can help to partially offset the higher development costs associated with these items.  **Technology-enhanced items (TEIs)** feature technology enhancements—such as the addition of interactive functionality, sound, graphics, animation, or video—to the stimulus materials, item stem, response options, and/or response mode, where the technology is intended to substantially improve some aspect of measurement. These items elicit discrete and brief responses compared to more continuous digital assessment approaches, such as [online games and simulated environments](https://neo.pearson.com/docs/DOC-197128).  TEIs are typically machine scorable, supporting either dichotomous or partial credit scoring models. As a result of these features, TEIs may expand the range of skills that are practically assessable in large-scale contexts compared to traditional paper-and-pencil formats. By virtue of their enhanced interactivity and functionality, TEIs may be more appealing and engaging to examinees than static, paper-based items. Finally, some types of TEIs may better support formative use relative to traditional paper-and-pencil items because they can be scored automatically and provide real-time performance feedback.  There is a substantial amount of overlap between SCR and TEI types: many SCR items, if administered online with additional features or functionality that enhance measurement quality in relation to what can be achieved via paper, would also be considered TEIs. However, SCR items are not necessarily technology-enhanced: SCRs are commonly administered via paper-and-pencil in the context of classroom or formative assessment. Similarly, not all TEIs call for constructed responses: some computer-administered selected-response items make use of technological enhancements to better measure complex reasoning or other higher-order thinking skills. Finally, not all TEIs and SCRs can be considered examples of [performance assessment](https://neo.pearson.com/docs/DOC-172914). For example, items that don’t meet the criterion of assessing complex reasoning or other higher-order thinking skills (a short-answer item or online version of a SR item that focuses on simple recall) would not be included within the framework. |
| Characteristic Features | Short Constructed-Response Items   * Whether in a paper-based or computer-administered environment, focus on production, development, or composition as opposed to selection from among pre-specified options * Elicit textual or graphical responses that are discrete and brief * Often allow multiple appropriate answers and/or partial credit scoring * Enable more direct measurement of certain types of skills than selected-response (SR) formats (e.g., graphing, editing) * May offer more realistic or authentic problem representations than SR items * May better support diagnosis and remediation of certain specific skill deficiencies than SR items * May be human or machine-scorable   Technology-Enhanced Items   * Offer technological enhancements—such as the addition of interactive functionality, sound, graphics, animation, or video—to the stimulus materials, item stem, response options, and/or response mode, where the technology substantially improves some aspect of measurement by measuring something better (more validly, reliably, or efficiently) or measuring more (increased range of content, knowledge, or skills) * Elicit responses that are discrete and brief compared to more continuous digital approaches, such as online games and simulated environments * Are typically machine scorable (either dichotomously or polytomously) * Expand the range of content, knowledge, or skills that are assessable in large-scale contexts compared to traditional paper-and-pencil formats, by either supporting new methods for capturing diverse response modes (e.g., to include speaking, listening, or presentation skills) or by making certain types of performances machine scorable that would have been hand-scored if administered via paper (e.g., creating a number line, graphing) * May be more appealing and engaging to examinees than static, paper-based items by virtue of their enhanced interactivity and functionality * May offer greater potential for formative assessment relative to traditional paper-and-pencil items vis-à-vis automated scoring and provision of real-time feedback |
| Design Variations and Other Considerations | **Short Constructed-Response Items**  SCR items can come in many different forms, depending on the type of response collected from the examinee—whether textual or graphical—including short-answer items, figural constructed response items, concept maps (Chung & Baker, 1997; Yin et al., 2005), and editing exercises (Breland, 1999; Davey, Godwin, & Mittelholtz, 1997), among others. Short-answer items might require examinees to compose a few words or a few sentences in response to a specific question, to solve a mathematical problem and provide a short explanation of the solution, or to identify elements in a series or list. Figural constructed-response items require examinees to complete, correct, or create drawings, illustrations, or graphics, such as graphs and charts. Concept maps require examinees to complete or construct a graphic representation of some topic by organizing and illustrating relationships between key terms. Editing exercises present examinees with a flawed writing sample and require them to make revisions to improve the writing.    There are other examples of SCR item formats, and within the formats described above, there are additional variations depending on the degree of constraint or open-endedness of the item. What all of these item formats have in common is that examinees are required to produce, develop, or construct their responses rather than select from among a limited set of pre-specified options. However, the level of production or development can vary across different item types, ranging from partial development (e.g., sentence completion or equation correction) to full development (construction of a total unit or response, as in a short answer item).  **Technology-Enhanced Items**  TEIs can also come in many different forms, depending on the response demand and response mode. Scalise and Gifford (2006) identified a continuum of TEI types, ranging from most to least constrained. Many of the item types in their continuum overlap with commonly-mentioned SCR item types, such as figural constructed response items and concept maps. However, TEIs may also include more constrained item types than are typically associated with SCR items, such as reordering or rearrangement items, where responses are selected from a pre-specified set, but examinees are required to order or sequence their responses in some way. As noted by Parshall et al. (2002), one of the benefits of TEIs compared to paper-based items is that TEIs can reduce the impact of random guessing by substantially increasing the number of possible response options available in more constrained TEI types.    What all of these item types have in common is that the item features some form of technology enhancement—in the form of embedded media or additional functionality—that enables better, more direct, or more complete measurement of the targeted skill. However, as with SCRs, the level of production or development called for can vary across different item types, ranging from partial development (e.g., editing a writing sample) to full development (construction of a total unit or response, as in graph construction).    TEIs can also be distinguished along other dimensions. For example, items may or may not incorporate some form of media, such as sound, animation, or video; and items can differ with respect to the types of associated functionalities they incorporate, such as the ability to rotate, resize, zoom, pause, or playback. Items can also differ with respect to the level of interactivity supported, or the extent to which the item responds or reacts to examinee inputs. Interactivity can range from “passive” interaction, in which the interface responds to examinee inputs by displaying in real-time the results of any changes or actions applied (e.g., a bar chart that allows examinees to change the height of bars by dragging them), to more interactive applications (e.g., multi-step, adaptive items or items that can provide immediate feedback based on the examinee’s response). Parshall et al. (2000) point out that, although increased interactivity improves the “realism” of the examinee’s experience compared to more static item types, it also increases the complexity of scoring. |
| Response Demands | **Short Constructed-Response Items**  The response demands of SCR items can vary depending on the level of constraint of the particular SCR format used. As previously explained, SCR items require examinees to develop, produce, or create a partial or full response to an item (as opposed to selecting from among a limited set of pre-specified response options). However, that response can take many forms, ranging from a few words or sentences to a graphic representation of a topic that illustrates the organization of and relationships among key concepts. Depending on the level of constraint of the item type used, SCRs may require examinees to complete or correct a set of stimulus materials (partial response) or to formulate a total idea unit (full response). Due to relaxed levels of constraint compared to traditional selected-response (SR) item types, SCRs are said to allow better or more direct measurement of hard-to-assess knowledge and skills, such as complex reasoning and other higher-order thinking skills (Bennett, 1993; Bennett et al., 1991).    **Technology-Enhanced Items**  Similar to SCRs, TEIs can demand a wide range of responses from examinees depending on the level of constraint of the particular TEI format used. As with SCRs, examinees may be required to develop, produce, or create a partial or full response to an item. In addition, although all TEIs require examinees to interact with a computer interface in some way, the range of possible response actions, or the means by which examinees can provide their responses (e.g., keyboard entry, mouse clicks, touch screens, or speech recognition software) is broad. Due to relaxed levels of constraint, as well as enhanced examinee interest and engagement compared to more static items, TEIs are said to allow better or more direct measurement of hard-to-assess knowledge and skills, such as complex reasoning and higher-order thinking skills (Parshall et al., 2000; Scalise & Gifford, 2006). In addition, the increased range of response actions may reduce the impact of construct-irrelevance associated with limited response modes by making items more accessible to all students (Dolan et al., 2011). |
| Evaluation Criteria and Procedures | **Short Constructed-Response Items**  SCR items are typically scored using rubrics that identify one or more appropriate or correct responses. Rubrics may also identify responses characterized by varying degrees of quality or correctness, thereby supporting partial credit scoring models. Scoring complexity varies depending on whether items are scored dichotomously or polytomously, as well as the degree of constraint of the item type. For example, open-ended items that allow partial credit scoring and require full or complete construction imply greater scoring complexity than more constrained items that require only partial construction and are scored dichotomously. Currently, some types of SCR items are machine-scorable using either rule-based scoring methods or algorithms based on natural language processing methods (Leacock, 2004).    **Technology-Enhanced Items**  Scoring method also varies across different TEI types. Less constrained TEI types typically require a scoring rubric identifying one or more appropriate or correct responses. In addition, as with SCR items, some TEIs are scored dichotomously, whereas others are scored polytomously, supporting partial credit scoring models. The degree of interactivity can also impact scoring complexity, both at the micro and macro assessment levels. For example, more interactive TEIs require more complex scoring rules, and assessments composed of greater numbers of TEIs require more complex scoring models that can handle large amounts of scoring dependency both across and within items (Parshall et al., 2000). TEIs are typically machine-scored using either rule-based scoring methods or algorithms based on natural language processing methods (Scalise & Gifford, 2006). |
| Administration Time | Short, brief; response times can range from one to several minutes |
| References | Bennett, R. E. (1993). On the meanings of constructed response. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment*(pp. 1-27). Hillsdale, NJ: Lawrence Erlbaum Associates.  Bennett, R. E., Sebrechts, M. M., & Yamamoto, K. (1991). *Fitting new measurement models to GRE General Test constructed-response item data*(RR-91-60). Princeton, NJ: Educational Testing Service.  Breland, H. M. (1999). *Exploration of an automated editing task as a GRE Writing measure*. (RR-99-9). Princeton, NJ: Educational Testing Service.  Chung, G. & Baker, E. (1997*). Year 1 technology studies: Implications for technology in assessment* (CSE Technical Report No. 459). Los Angeles: Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, University of California, Los Angeles.  Davey, T., Godwin, J., & Mittelholz, D. (1997). Developing and scoring an innovative computerized writing assessment. *Journal of Educational Measurement*, *34*, 21–41.  Dolan, R. P., Goodman, J., Strain-Seymour, E., Adams, J., & Sethuraman, S. (2011).*Cognitive lab evaluation of innovative items in mathematics and English/ language arts assessment of elementary, middle, and high school students*. Research Report. Iowa City: Pearson.  Leacock, C. (2004). Scoring free-responses automatically: A case study of a large-scale assessment. *Examens*, 1 (3).  Osterlind, S. J. (1998). *Constructing test items: Multiple-choice, constructed-response, performance, and other formats*. Norwell, MA: Kluwer Academic Publisher.  Parshall, C. G., Davey, T., & Pashley, P. J. (2000). Innovative item types for computerized testing. In W. Van der Linden, Glas, C. A. W. (Ed.),*Computerized adaptive testing: Theory and practice* (pp. 129–148). Norwell, MA: Kluwer Academic Publisher.  Parshall, C. G., Spray, J., Kalohn, J., & Davey, T. (2002). Issues in innovative item types. In *Practical considerations in computer-based testing* (pp. 70–91). New York: Springer.  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).  Yin, Y., Vanides, J., Ruiz-Primo, M. A., Ayala, C. C., & Shavelson, R. J.  (2005). Comparison of two concept-mapping techniques: Implications for scoring, interpretation, and use. *Journal of Research in Science Teaching*, *42*(2): 166-184. |

## 04b Exemplars

|  |  |
| --- | --- |
| Page Title in Header | Examples of Short Constructed-Response and Technology-Enhanced Items |
| Header Text | Below is a list of innovative examples that demonstrate some of the ways that short constructed-response and technology-enhanced items may be used as performance assessment. |
| Body Header | [Smarter Balanced Assessment Consortium - Sample ELA TEI Items](http://sampleitems.smarterbalanced.org/itempreview/sbac/ELA.htm) |
| Body Copy | The Smarter Balanced Assessment Consortium (SBAC) has published English Language Arts technology-enhanced items (TEI) prototypes that demonstrate a variety of technical tools that can be used to measure student knowledge, skills, and abilities. To view these, use the filter tool on the SBAC site to identify the items they have categorized TEIs.  SBAC. Retrieved February 27, 2013from <http://www.smarterbalanced.org/> |
| Body Header | [DCD Developed TEIs and Short Constructed-Response Items](https://test.testnav.com/refqc/testnav-7.5.12.58/epatLogin.jsp?testnavTestId=cssmini&testnavFormId=cssmini) |
| Body Copy | Pearson's Digital Content Development team has developed a number of highly innovative tools and features that are used for technology-enhanced and other item types delivered online within the TestNav™ online delivery system. This link provides two examples: one using a graphing tool and providing space for a constructed response and another that allows students to create a presentation within an online form.  Note the sample items will be launched in TestNav using a non-secure practice tool. To run TestNav, your machine must meet the minimum requirements found here: <http://www.pearsononlinetesting.com/TestNav/7/requirements_testnav_7_5_10.html> |

# 05 Essays

|  |  |
| --- | --- |
| Page Title in Header | Essays |
| Header Text | Essays require examinees to compose an extended piece of writing in response to a specific prompt. |
| Body Header |  |
| Body Copy | In general, features of essays include:   * Requiring students to compose an extended piece of writing (in contrast to short constructed-response items) * Posing a specific writing prompt that often identifies the purpose for writing, the intended audience, and desired structural or content-related features of the response * Providing a more direct measure of writing skills than selected-response formats * Scoring using one or more rubrics, which may be analytic or holistic * Human or machine scoring |
| Brainshark | Yes |

## 05a More Info

|  |  |
| --- | --- |
| Page Title | What Are Essays? |
| Definition | Essay prompts require students to compose an extended piece of writing in response to a specific writing prompt. Such responses are lengthier than [short constructed responses](https://neo.pearson.com/docs/DOC-208654), capable of treating a given topic or purpose in more depth and providing students more opportunities to display composition skills. Essay prompts often identify the purpose for writing, the intended audience, and/or one or more desired structural or content-related features of the response. By virtue of their emphasis on composition, production, or development (as opposed to selection among alternatives), essays provide a more direct measure of a wider range of writing skills and processes compared to more constrained assessment approaches (selected-response items focusing on editing subskills, for example).  In general, when used in large-scale summative assessment contexts (e.g., national or state writing assessment programs), essays constitute discrete writing samples, under timed and relatively constrained administration conditions, with typically no opportunities for examinees to create successive drafts, receive feedback, and make revisions. Although these conditions are contrary to good writing practices, strict standardization of assessment conditions is important for maintaining comparability and score quality in summative contexts. When used to support formative assessment, on the other hand, essays can more closely model the stages of an authentic writing process (e.g., including prewriting/planning, drafting, peer reviewing, and revising) and can be distributed over multiple occasions. Essays are either human or machine-scored using one or more rubrics or sets of scoring criteria that recognize varying gradations of response quality, as well as attributes or features that characterize each level. Rubrics may require an overall judgment of response quality (holistic) or require separate judgments of response quality along multiple dimensions or traits (analytic). |
| **Common Contexts** | Essays can support a variety of writing purposes. Examples of different types of prompts for different purposes include:   * Explaining or analyzing via reading to write prompts (e.g., “Summarize the main idea presented in the text…”) * Telling a story through a personal narrative (e.g., “Tell a true story about a time you overcame a challenge…”) * Persuading others through argument-based prompts (e.g., “Use data and evidence from the text to evaluate the claim that…”) * Representing a scene or event using sensory details through descriptive prompts (e.g., “describe your favorite place so that someone reading your essay could imagine it”) |
| Design Variations and Other Considerations | Essay prompts can vary along a number of dimensions. First, prompts may focus on assessing writing skills, mastery of subject-matter or content knowledge, or both. If both writing skills and content knowledge are assessed, distinct rubrics may be developed to produce separate scores—one for the content being assessed and one for the quality of the writing.  Prompts may also vary with respect to whether they only require students to produce writing (writing only prompts) or whether they also require students to read or otherwise synthesize and integrate a set of stimulus materials (e.g., passages, charts, graphs, or images) into their written responses (reading to write prompts).  Some assessment programs allow students to choose from among several topics, whereas others prescribe the use of a specific prompt for all students in order to ensure comparability of scores. Research on the impact of student choice on performance has been mixed, with some evidence suggesting that giving students choices allows them to demonstrate their maximal performance but does not alter the meaning of scores (Bridgeman, Morgan, & Wang, 1997), and other evidence suggesting that examinees are not equally skilled in selecting the “best” prompt, resulting in scores that are not comparable (Linn, Betebenner, & Wheeler, 1998; Wainer & Thissen, 1994).  Essay prompts also vary by discourse mode (Huot, 1990). Narrative writing prompts, which can be personal or imaginative, require students to recount an event or tell a story; descriptive writing prompts ask examinees to use sensory details to depict a person, place, or thing; expository writing provides information about or an explanation of an issue or topic; and argument-based essays ask writers to use evidence or reason to evaluate the veracity of a claim.  Prompts may also differ with respect to the level of rhetorical specification provided (Huot, 1990). Prompts may range from relatively open-ended (identifying only the topic, for example) to highly specified (identifying topic, purpose for writing, intended audience, and context). More open-ended prompts may be more difficult to score reliably than highly specified prompts, due to greater variability in how students approach the task. In addition, some writing proponents claim that highly specified prompts are more authentic to the types of writing students will perform in future settings (e.g., in professional contexts). However, in high-stakes writing assessments, even open-ended prompts implicitly specify both a purpose (obtain the highest score possible) and an audience (trained rater). Moreover, the addition of lengthy contexts may not improve writing performance and may introduce construct-irrelevant variance through unnecessary reading load.  Prompts may also vary with respect to the amount of structure provided to support student responses. For example, some prompts may simply specify what the final writing product should look like, whereas others may direct students to brainstorm ideas, draft, and revise. Some essay prompts may even provide material supports for more extended writing processes, including graphic organizers to help students prepare and plan their responses.  Variations in wording of essay prompts are also possible (e.g, personal versus neutral, question versus command). Research evidence is mixed as to whether minor variations in prompt wording have an effect on student performance (Huot, 1990). Essays may be delivered on paper or computer-administered. Research generally suggests comparability of scores from online and paper-based administrations of essays, assuming any handwritten responses are typed or transformed using a word processing program prior to scoring (Bennett, 2003).  Finally, essay prompts can support both formative and summative uses. In formative settings, the essay approach may more closely resemble an authentic writing process, with writing distributed over several occasions and opportunities for students to plan, create successive drafts, receive feedback, and revise their writing. In most large-scale summative settings, where strict standardization is more important to ensure comparability, writing may be limited to a single occasion, students typically do not receive interim feedback on drafts, and the prompts themselves (e.g., topics, discourse mode, level of specification, wording) are often more carefully controlled. There are exceptions to this rule, however. A few state or district writing assessment programs provide lengthier testing windows, multiple writing occasions, and opportunities for students to draft and revise. However, none of these programs provides the option for students to receive teacher or peer feedback on their drafts. |
| Response Demands | All essays require students to compose a piece of writing on demand. Even when pre-writing planning activities are not explicitly required or scored, planning may be necessary in order to complete a response that addresses the prompt within the time limits. However, planning behavior probably varies as a function of writing ability, with stronger writers presumably engaging in more planning than weak writers.    Response demands may vary depending on the type of prompt (whether writing only or reading to write). With the former type of prompt, students are only required to read the prompt itself and their responses are scored only on the basis of their writing skills. With the latter type, students may also have to read, comprehend, and integrate information from associated stimulus materials, such as passages, graphs, or charts. Further, their responses may be evaluated with respect to mastery of specific subject-matter content in addition to quality of the writing. Different discourse modes may also place different response demands on students (Huot, 1990; Quellmalz et al., 1980). Students creating an imaginative narrative, for example, will probably use different writing strategies than students creating an argument. Research demonstrates that student performance tends to vary across different discourse modes, suggesting the need to sample from different types of writing in order to make valid inferences about general writing ability (Huot, 1990; Quellmalz et al., 1980).    Finally, response demands may differ by examinee subgroup. Research shows persistent subgroup differences in essay performance by gender (Willingham et al., 1997), ethnicity (Applebee & Langer, 2006), and native language status (Heck & Crislip, 2001), with the most recent NAEP results exhibiting all three types of differences. These disparities in performance may indicate real differences in writing ability rather than bias. However, they may also indicate differences in the writing strategies used by different types of examinees. More research is needed in this area. |
| Evaluation Criteria and Procedures | Essay responses are generally scored using one or more rubrics. Holistic rubrics provide an overall judgment of response quality, whereas analytic rubrics provide separate judgments of response quality along multiple dimensions or traits. Although the writing community is not in complete agreement on the qualities of “good” writing, commonly assessed traits include content, organization, development, voice, vocabulary, language, mechanics, and coherence.  There is a large body of research on the relative advantages and disadvantages of analytic and holistic rubrics. This body of research suggests some general conclusions: holistic scoring is faster and less expensive than analytic scoring, but analytic scores may be more reliable; scores assigned using analytic and holistic rubrics tend to be very highly correlated; only a few writing traits can be reliably distinguished using an analytic rubric; and analytic rubrics better support instruction and learning than holistic rubrics because they provide more information regarding specific strengths and weaknesses (Huot, 1990; Quellmalz, 1984). However, research also suggests that when multiple analytic scores are assigned by the same rater (which is a common scoring method due to cost efficiency), there is the potential for halo error. In other words, raters’ perceptions of one trait may affect their perceptions of other traits, leading to spuriously high correlations between separate traits (Lai, Wolfe, & Vickers, 2012).  Essays may be human- or machine-scored. Research on automated scoring systems demonstrates high agreement between automated scores and scores assigned by human raters (Attali, 2007; Landauer, Laham, & Foltz, 2003; Nichols, 2004; Page, 2003). There are currently a number of different automated scoring approaches that differ from one another primarily with respect to their underlying scoring algorithms. However, they all attempt to predict human scores by tagging response features that characterize different levels of quality and aggregating those features into some sort of meaningful composite. Different scoring models recognize different response features and may aggregate those features differently. Scoring engines are trained and calibrated with human-scored responses. Therefore, the quality of the human-scored calibration set (e.g., representativeness, accuracy and consistency of scores) is a crucial contributor to the quality of automated scores. |
| Administration Time | Most large-scale essay prompts range from 25 to 90 minutes. In contrast, essays administered in the classroom to support formative use might be distributed over several days. |
| References | Applebee, A. N. & Langer, J. A. (2006). *The state of writing instruction in America’s schools: What existing data   tell us*. Albany, NY: Center on English Learning and Achievement.  Attali, Y. (2004). *Construct validity of e-rater in scoring TOEFL essays* (Research Report). Princeton, NJ: Educational Testing Service.  Bennett, R. E. (2003). *Online assessment and the comparability of score meaning*. Princeton, NJ: Educational Testing Service.  Bridgeman, B., Morgan, R., & Wang, M. M. (1997). Choice among essay topics: Impact on performance and validity. *Journal of Educational Measurement, 34*(3): 273-286.  Chung, G. K. W. K. & O’Neil, H. F. (1997). *Methodological approaches to online scoring of essays* (CSE Technical Report No. 461). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.  Crusan, D. (2002). An assessment of ESL writing placement assessment. *Assessing Writing, 8*, 17-30.  Heck, R. H. & Crislip, M. (2001). Direct and indirect writing assessments: Examining issues of equity and utility. *Educational Evaluation and Policy Analysis, 23*(1): 19-36.  Huot, B. (1990). The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research, 60*(2): 237-263.  Lai, E. R., Wolfe, E. W., & Vickers, D. H. (2012). *Halo effects and analytic scoring: A summary of two empirical studies* (Research Report). Iowa City, IA: Pearson.  Landauer, T. K., Laham, D., & Foltz, P. (2003). Automatic essay assessment. *Assessment in Education, 10*(3): 295-308.  Linn, R. L., Betebenner, D. W., & Wheeler, K. S. (1998). *Problem choice by test-takers: Implications for comparability and construct validity* (CSE Technical Report No. 495). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.  National Assessment Governing Board. (2010). *Writing framework for the 2011 National Assessment of Educational Progress*. Washington, D.C.  Nichols, P. D. (2004, April). Evidence for the interpretation and use of scores from an automated essay scorer. Paper presented at the Annual Meeting of the American Educational Research Association (AERA), San Diego, CA.  Page, E. B. (2003). Project essay grade: PEG. In M. D. Shermis & J. Burstein (Eds.),*Automated essay scoring: A cross-disciplinary perspective* (pp. 43–54). Mahwah, NJ: Lawrence Erlbaum Associates.  Quellmalz, E. S. (1984). Designing writing assessments: Balancing fairness, utility, and cost. *Educational Evaluation and Policy Analysis, 6*(1): 63-72.  Quellmalz, E., Capell, F. J., & Chou, C. P. (1980). *Defining writing: Effects of discourse and response mode* (CSE Report No. 132). Center for the Study of Evaluation: Los Angeles.  Stiggins, R. J. (1982). A comparison of direct and indirect writing assessment methods.*Research in the Teaching of English, 16*(2): 101-114.  Wainer, H. & Thissen, D. (1994). On examinee choice in educational testing. *Review of Educational Research*, 64(1): 159-195.  Willingham, W. W., & Cole, N. S. (1997). *Gender and fair assessment*. Mahwah, NJ:  Erlbaum. |

## 05b Exemplars

|  |  |
| --- | --- |
| Page Title in Header | Examples of Essays |
| Header Text | Below is a list of innovative examples that demonstrate some of the ways that essays may be used as performance assessment. |
| Body Header | [National Assessment of Educational Progress](http://nces.ed.gov/nationsreportcard/) |
| Body Copy | The National Assessment of Educational Progress (NAEP) has released several writing prompts and rubrics from previous administrations.   * Released writing prompts (grades 4, 8, and 12): <http://nces.ed.gov/nationsreportcard/itmrlsx/search.aspx?subject=writing> * Rubrics: [http://www.nagb.org/content/nagb/assets/documents/publications/frameworks/writin g-2011.pdf](http://www.nagb.org/content/nagb/assets/documents/publications/frameworks/writing-2011.pdf)   NAEP. Retrieved March 14, 2013 from http://nces.ed.gov/nationsreportcard/ |
| Body Header | [Oklahoma State Department of Education, Oklahoma Core Curriculum Tests](http://ok.gov/sde/accountability-assessments) |
| Body Text | The Oklahoma State Department of Education (ODE) includes writing in their Oklahoma Core Curriculum Tests (OCCT) for students in grade 5. As an example of what might appear on the assessment, the ODE has provided the public with a narrative writing prompt and rubric.  Grade 5 narrative writing prompt: <http://ok.gov/sde/sites/ok.gov.sde/files/Gr5_Narrative_Prompt.pdf>  Rubric: <http://ok.gov/sde/sites/ok.gov.sde/files/CCSS_Gr5_Trans_Rubric.pdf>  ODE. Retrieved March 14, 2013 from [http://ode.gov/sde](http://ok.gov/sde/) |

# 06 Performance Tasks

|  |  |
| --- | --- |
| Page Title in Header | Performance Tasks |
| Header Text | Performance tasks require examinees to use one or more stimulus materials in responding to a series of discrete items or activities that are aligned around a common theme or culminating activity. |
| Body Header |  |
| Body Copy | In general, features of performance tasks include:   * Items or activities organized around a theme or culminating activity * Contexts that simulate realistic instructional, professional, or on-the-job activities * A focus on valued learning outcomes * Alignment to multiple content standards, skills, or processes * Evaluation procedures that typically involve the application of one or more rubrics |
| Brainshark | Yes |

## 06a More Info

|  |  |
| --- | --- |
| Page Title | What Are Performance Tasks? |
| Definition | Performance tasks require examinees to process and use one or more stimulus materials in responding to a series of discrete items or activities that are organized coherently around a common theme or culminating activity. Collectively, these activities reflect multiple content standards, skills, and/or processes that represent valued learning outcomes. Performance tasks are typically embedded in contexts that simulate realistic instructional, professional, or on-the-job activities, where the simulation is characterized by at least a moderate degree of fidelity between the assessment activities and the criterion performance(s) to which we would like to generalize. One or more component items within a performance task are typically evaluated using a rubric. |
| Characteristic Features | * Collection of discrete items or activities organized around a theme or culminating activity * Makes use of one or more stimulus materials * Set in contexts that simulate realistic instructional, professional, or on-the-job activities * At least a moderate degree of fidelity between the assessment activity and the criterion performance * Focus on valued learning outcomes * Align to multiple content standards, skills, or processes * Evaluation procedures typically involve the application of one or more rubrics |
| Common Contexts | Performance tasks are commonly used in classroom formative contexts (e.g., a biology lab exercise), summative K-12 settings (e.g., a district/state assessment or an Advanced Placement exam), and certification or licensure examinations. |
| Design Variations and Other Considerations | **Item and Scoring Options**  The discrete items or activities that comprise a task may combine several formats, including selected-response, technology-enhanced, or constructed-response (short or extended). These items or activities may or may not culminate in an extended performance that occurs at the end of the task and ties together preceding activities.  Depending on the types of items or activities included in a task, items may be machine-scored, human-scored, or some combination. Machine scorable items may include those scored with an automated scoring engine.    **Purposes**  Performance tasks can be used for summative purposes (e.g., holding teachers or students accountable for learning) or for classroom formative purposes. Performance tasks designed to fulfill one of these goals will not necessarily collect information useful for serving the other. For example, tasks used for high-stakes summative purposes require a higher level of standardization of task administration conditions (e.g., testing time, student instructions, prompts, available resources and supports) than tasks used for classroom formative purposes. In the latter context, tasks may be modified to fit student interest or ability by allowing students to choose the manner in which they will respond or by allowing students to submit portions of the task over an extended period of time. |
| Response Demands | In a performance task, examinees are asked to respond to the discrete items or activities that comprise the task. Depending on the purpose of the task, students may respond orally, in writing, via a concrete work product (e.g., a graph or drawing), or some combination of the three. For performance tasks that include technology-enhanced items, students respond using technology capabilities, such as clicking, dragging, or otherwise manipulating the online interface.    What distinguishes performance tasks from other approaches of performance assessments is the wide range of potential response demands and modes that can be represented within a single task. |
| Evaluation Criteria and Procedures | Performance tasks (or pieces of them) commonly require application of one or more rubrics. Typically, there are separate rubrics created for each discrete item or activity within a task, but one might also create a rubric for the overall task performance. Rubrics may be holistic or analytic. They may focus on a single dimension of performance or encompass multiple dimensions of performance, depending on the purpose of the assessment. For example, performance tasks that feature a culminating written performance might focus on both the content of the essay (e.g., student knowledge of a specific history topic), as well as the quality of the writing (e.g., organization, voice, use of writing conventions). |
| Administration Time | May range from 45 minutes to several hours or even over several days or weeks |
| References | Landauer, T. K., Laham, D., & Foltz, P. (2003). Automatic essay assessment. *Assessment in Education,* 10(3): 295-308.  Nichols, P. D. (2004, April). Evidence for the interpretation and use of scores from an automated essay scorer. Paper presented at the Annual Meeting of the American Educational Research Association (AERA), San Diego, CA.  Osterlind, S. J. (1998). C*onstructing test items: Multiple-choice, constructed-response, performance, and other formats.* Norwell, MA: Kluwer Academic Publisher.  Partnership for Assessment of Readiness for College and Careers. (2010). Application for the Race to the Top Comprehensive Assessment Systems Competition. Retrieved from http://www.fldoe.org/parcc/pdf/apprtcasc.pdf.  SMARTER Balanced Assessment Consortium. (2010). Race to the Top assessment program application for new grants. Retrieved from http://www.k12.wa.us/SMARTER/RTTTApplication.aspx.  Stiggins, R. J. (1987). The design and development of performance  assessments. *Educational Measurement: Issues and Practice,* 6, 33-42.  Wiggins, G. (1993). *Assessing student performance.* San Francisco: Jossey-Bass Publishers |

## 06b Exemplars

|  |  |
| --- | --- |
| Page Title in Header | Examples of Performance Task |
| Header Text | Below is a list of innovative examples that demonstrate some of the ways that performance tasks may be used as performance assessment. |
| Body Header | [Partnership for the Assessment of Readiness for College and Careers (PARCC) - Sample Grade 10 Literary Task](http://www.parcconline.org/samples/english-language-artsliteracy/grade-10-elaliteracy) |
| Body Copy | The Partnership for Assessment of Readiness for College and Careers (PARCC) has released several prototype performance tasks, such as this grade 10 literary task, which requires students to carefully read two complex texts, respond to evidence-based selected-response items, and then compose two analytic essays.PARCC. Retrieved November 27, 2012 from [http://www.parcconline.org](http://www.parcconline.org/) |
| Body Header | [New York City Department of Education (NYC DOE)](http://schools.nyc.gov/NR/rdonlyres/84642811-B9DA-47E5-BC80-92925326B2C9/0/NYCDOE_G2_LiteracySS_WheresHome_Final.pdf) |
| Body Copy | NYC educators and national experts have developed Common Core-aligned tasks, such as this example from grade 2. where students read informational texts, participate in the creation of a whole class essay about one type of community, , and work independently on essays about the two others.  NYC DOE. Retrieved November 27, 2012 from [http://schools.nyc.gov](http://schools.nyc.gov/) |

# 07 Demonstrations

|  |  |
| --- | --- |
| Page Title in Header | Demonstrations |
| Header Text | Demonstrations require examinees to engage in a live performance of the targeted knowledge, skills, or abilities (KSAs), where the assessment conditions resemble real-life situations. |
| Body Header |  |
| Body Copy | In general, features of the demonstrations include:   * A live performance * Evaluation that focused on both process and product (in contrast to performance tasks) * The input by subject-matter or professional experts on the scoring criteria * A high degree of similarity between assessment conditions and conditions of real performance experienced in a particular discipline or domain |
| Brainshark | Yes |

## 07a More Info

|  |  |
| --- | --- |
| Page Title | What are demonstrations? |
| Definition | Demonstrations require examinees to engage in a live performance of the targeted knowledge, skills, or abilities (KSAs), where the assessment conditions resemble to a high degree of fidelity in construct-relevant ways the conditions under which the targeted KSAs are typically exercised in a given domain or discipline. Such performances must be scored by trained raters, and evaluation of performance quality tends to focus on both process and work products. |
| Characteristic Features | * Live performance * Evaluation focuses on both process and product (in contrast to performance tasks) * Development of evaluation criteria requires input by subject-matter or professional experts * High fidelity, authenticity, or degree of similarity on construct-relevant dimensions between assessment conditions and conditions of real performance experienced in a particular discipline or domain |
| Common Contexts |  |
| Design Variations and Other Considerations | **Scoring Expertise**  Although experts agree that evaluation of demonstrations requires the judgments of experts, it is unclear whether this implies that scorers must have domain or content expertise or merely scoring expertise. Regardless of who performs the scoring, however, input from domain experts will always be necessary in some form or another during task construction and the design of the evaluation criteria. |
| Response Demands | In a demonstration, examinees are asked to do, perform, create, or construct something. The particular response demands of a given demonstration depend on the domain of interest and the particular assessment context.    What distinguishes demonstrations from other performance assessment approaches is that the response demands in the assessment context are similar to the response demands of the domain or discipline to which we would like to generalize. For example, in an assessment certifying physicians’ diagnostic skills, examinees are expected to interact with patients, ask questions, and interpret patient responses. These types of responses are highly similar to, if not the same as, the response demands of a practicing physician. Similarly, in a demonstration of students’ oral reading fluency, students are asked to read short passages aloud. Thus, the response demand of the assessment matches the response demand of the domain (i.e., oral reading ability) about which we want to make inferences. |
| Evaluation Criteria and Procedures | The criteria for evaluating performances are based on both the primary focus of the observation and the goals of the assessment. Performances are commonly scored using rubrics or observational checklists to be completed by trained raters. The range of performance quality represented in these rubrics and checklists ought to be identified in collaboration with domain or discipline experts and reflect the intent of the standards within a domain or discipline. |
| Administration Time | Varies from a few minutes (e.g., pull-ups) to 20 minutes (e.g., a dance) or more (e.g., sporting event) |
| References | Baxter, G. P., Shavelson, R. J., Goldman, S. R., & Pine, J. (1992). Evaluation of procedure-based scoring for hands-on science assessment. *Journal of Educational Measurement*, 29, 1-17.  Behrens, J. T., DiCerbo, K. E., & Ferrara, S. (2012). Intended and unintended deceptions in the use of    simulations. Paper presented at the Technology Enhanced Assessment Symposium. Washington, D.C.  Bennett, R. E. (1993). On the meanings of constructed response. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment*(pp. 1-27). Hillsdale, NJ: Lawrence Erlbaum Associates.  Gitomer, D. H. (1993). Performance assessment and educational measurement. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment*(pp. 241-263). Hillsdale, NJ: Lawrence Erlbaum Associates.  Lane, S. (2010). *Performance assessment: The state of the art*. (SCOPE Student Performance Assessment Series). Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.  Martin, J. A., Reznick, R. K., Rothman, A., Tamblyn, R. M., & Regehr, G. (1996). Who should rate  candidates in an objective structured clinical examination?*Academic Medicine*, 71, 170–175. |

## 07b Exemplars

|  |  |
| --- | --- |
| Page Title in Header | Examples of Demonstrations |
| Header Text | Below is a list of innovative examples that show some of the ways that demonstrations may be used as performance assessment. |
| Body Header | [VETASSESS Practical Trade Skill Assessment – Plumbing, Journeyman Examination](http://www.vetassess.com.au/take_a_test/plumbing.cfm) |
| Body Copy | This assessment program requires workers to demonstrate a specific set of job functions, such as might be required of a plumber, to earn a trade certificate or course credit based on their demonstrated skills.    Vetassess. Retrieved November 27, 2012 from <http://www.vetassess.com.au/> |
| Body Header | [Stanford Education Assessment Laboratory (SEAL)](http://www.stanford.edu/dept/SUSE/SEAL/) |
| Body Copy | **Click: Assessments/Instruments. Click: Paper Towels**  SEAL conducts assessment research that is both conceptual and methodological (psychometric, statistical, qualitative). This demonstration, requiring students to design and conduct an experiment, is an assessment of scientific reasoning, critical thinking, and inquiry.  Stanford Education Assessment Laboratory. Retrieved November 27,2012 from <http://www.stanford.edu/dept/SUSE/SEAL/> |

# 08 Projects

|  |  |
| --- | --- |
| Page Title in Header | Projects |
| Header Text | Projects are long-term, extended performances that are organized around a driving question or problem and often result in a culminating product or presentation. |
| Body Header |  |
| Body Copy | In general, features of projects include:   * Blending instruction, learning, and assessment * Opportunities for self-directed learning, student autonomy, and student decision-marking * Opportunities for student collaboration or consultation with others (expert, teacher, parent) * Require a diverse array of knowledge and skills |
| Brainshark | Yes |

## 07a More Info

|  |  |
| --- | --- |
| Page Title | What Are Projects? |
| Definition | Projects are long-term, extended performances that are organized around a driving question or problem and often result in a culminating product or presentation. Projects engage learners (alone or with others) in complex, multistep processes that require a diverse array of knowledge and skills: subject-matter knowledge (possibly from multiple disciplines), discrete skills, and higher-order competencies, such as problem solving, creativity, communication, collaboration, and metacognition.  Projects are typically framed within engaging, meaningful, and realistic contexts and provide opportunities for self-directed student learning and autonomy in producing the product or preparing the presentation. Projects represent an almost equal blending of instruction, learning, and assessment. |
| Characteristic Features | * Blend instruction, learning, and assessment * Are organized around a driving question or problem, often of a particularly complex or ill-structured nature * Are framed within meaningful, realistic contexts * Often result in a culminating product or presentation * Provide opportunities for self-directed learning, student autonomy, and student decision-making * May provide opportunities for student collaboration or consultation with others (e.g., expert, teacher, parent) * Require a diverse array of knowledge and skills * Require long-term, multi-step processes, potentially lasting several weeks or months * May require application of knowledge and skills from multiple disciplines * Believed to increase student interest, engagement, and motivation |
| Common Contexts | Projects are commonly used in K-12 classroom contexts (e.g., term papers and science fair projects) and postsecondary assessments (e.g., honors theses and group research presentations). |
| Design Variations and Other Considerations | All projects involve some degree of student self-directed learning, although the level of student autonomy supported within a given project can vary. For example, teachers may involve students in identifying the problem or driving question. Alternatively, the teacher may select the driving question, but allow students to choose their own methods for investigation or identify specific sub-topics to pursue.  Similarly, the teacher’s role during a project is to act as a facilitator, rather than as a leader or expert; however, the particular level of scaffolding, support, or direction a teacher provides may vary, depending on characteristics of the students or the nature of the project itself. The level of open-endedness of the problem or driving question can also vary, ranging from a problem with multiple possible solutions or multiple paths to a given solution to an indeterminate problem, or a problem with no known solution.    Projects may or may not involve collaboration with peers or consultation with others, such as teachers, parents, or experts. Examples of projects involving collaboration with others include group research projects, science fair projects, and presentations. There are also examples of projects that are completed primarily on an individual basis, such as a Master’s thesis or high school social studies research paper. Projects may integrate content and themes across multiple subjects, domains, or disciplines, or they may be focused within a single area of study. |
| Response Demands | In a project, examinees are asked to perform some combination of the following: articulate a problem, identify hypotheses, gather information, plan, self-monitor progress, work with others (communicate effectively, share responsibility), design and implement investigations, summarize results, and report findings via multiple communication modes, including written, graphical/visual, and oral. |
| Evaluation Criteria and Procedures | Projects offer opportunities for both formative and summative evaluation of student learning. Formative evaluation might focus on observations of student interactions with teammates or on interim artifacts, such as outlines, research plans, or reference lists that are produced over the duration of the project and form the basis for progress monitoring and feedback to students. Summative evaluation might focus on a culminating work product, such as a written report or presentation that is used to assign student grades.    Projects may be used to assess a wide range of student learning outcomes, ranging from traditional command of subject-matter knowledge to mastery of discrete skills and procedures, to more complex, so-called “21st century” skills: communication, collaboration, critical thinking and problem-solving, creativity, and self-directed learning or metacognition.    When projects are completed by groups rather than individuals, there are particular challenges associated with evaluation and scoring. For example, educators are typically interested in obtaining individual student grades, but group projects tend to obscure individual contributions (Webb, 1995). Often teachers will assign a single grade to a group based on completion of the group project, and this group score in turn is assigned to each individual group member. However, to the extent that individual student contributions are not comparable or equivalent, this does not necessarily reflect individual effort, knowledge, or skills (Race, 2001; Saner et al., 1994; Webb, 1995).  Research shows that group-level assessments may not yield scores that are predictive of individual-level ability, even when individual students turn in separate work products. In particular, scores from group work tend to over-estimate individual performance, and exhibit both ceiling effects and range restriction (Webb, 1993). If this higher performance reflects real learning, then group project scores are not necessarily invalid. However, if students obtain higher scores from group projects simply because their more able group mates complete most of the work, then using group scores as indicators of individual student learning is problematic. |
| Administration Time | Long-term or extended performances, ranging from several weeks to several months |
| References | Blumenfeld, P. C., Soloway, E., Marx, R. W., Krajcik, J. S., Guzdial, M., & Palinscar, A. (1991). Motivating project-based learning: Sustaining the doing, supporting the learning. Educational Psychologist, 26 (3/4), 369-398.  Donnelly, R. & Fitzmaurice, M. (2005). Collaborative project-based learning and problem-based learning in higher education: A consideration of tutor and student roles in learner-focused strategies. In O’Neill, G., Moore, S., McMullin, B. (Eds.), Emerging issues in the practice of university learning and teaching. Dublin: AISHE.  Doppelt, Y. (2003). Implementation and assessment of project-based learning in a flexible environment. International Journal of Technology and Design Education, 13, 255-272.  Hmelo-Silver, C. (2004). Problem-based learning: What and how do students learn?Educational Psychology Review, 16 (3), 235-266.  Major, C. H. & Palmer, B. (2001). Assessing the effectiveness of problem-based learning in higher education: Lessons from the literature. Academic Exchange Quarterly, 5 (1).  Race, P. (2001). A briefing on self, peer, and group assessment. Assessment Series Number 9. York, UK; Learning and Teaching Support Network.  Saner, H., McCaffrey, D., Stecher, B., Klein, S., & Bell, R. (1994). The effects of working in pairs in science performance assessments. Educational Assessment, 2 (4), 325–338.  Thomas, J. W. (2000). A review of research on project-based learning. The Autodesk Foundation: San Rafael, CA. Downloaded from [http://www.bie.org/research/study/ summary\_of\_research\_on\_project\_based\_learning](http://www.bie.org/research/study/summary_of_research_on_project_based_learning).  Webb, N. M. (1993). Collaborative group versus individual assessment in mathematics: Processes and outcomes. Educational Assessment, 1 (2), 131–152.  Webb, N. M. (1995). Group collaboration in assessment: Multiple objectives, processes, and outcomes. Educational Evaluation and Policy Analysis, 17 (2), 239–261. |

## 07b Exemplars

|  |  |
| --- | --- |
| Page Title in Header | Examples of projects |
| Header Text | Below is a list of innovative examples that demonstrate some of the ways that projects may be used as performance assessment. |
| Body Header | [Cornell University: BirdSleuth - Investigating Evidence](http://www.birds.cornell.edu/birdsleuth) |
| Body Copy | Through BirdSleuth, students learn to observe birds carefully, to ask questions based on their observations, and then to figure out how to answer their questions and share their results.  The Cornell Lab of Ornithology. Retrieved November 27, 2012 from <http://www.birds.cornell.edu/birdsleuth> |
| Body Header | [Buck Institute for Education: Gender Roles](http://www.bie.org/videos/video/learning_to_act_as_women_and_men) |
| Body Copy | Middle school students at the Greater Brunswick Charter School in New Jersey complete an interdisciplinary project about gender roles combining Language Arts, History, Biology, and Media Literacy.  Buck Institute for Education. Retrieved November 27, 2012 from <http://www.bie.org/videos/video/learning_to_act_as_women_and_men> |
| Body Header | [PBS: THE BLUES™ NATIONAL EDUCATIONAL OUTREACH CAMPAIGN](http://www.pbs.org/theblues/classroom.html) |
| Body Copy | The Blues educational outreach campaign, created by Experience Music Project, provides teaching strategies and resources for teachers to assist in the integration of blues music, culture, and history into American classrooms, with emphasis on grades 9-12.  PBS. Retrieved November 27, 2012 from <http://www.pbs.org/theblues/classroom.html> |
| Body Header | [Society of Workforce Planning Professionals: Certification Project](http://www.swpp.org/wmcertification.html) |
| Body Copy | This workforce management certification is composed of four parts, which include three tests and a project. The project is designed to be a “real-life” example of workforce management skills to demonstrate the application of knowledge.    SWPP. Retrieved November 27, 2012 from <http://www.swpp.org/wmcertification.html> |

# 09 Portfolios

|  |  |
| --- | --- |
| Page Title in Header | Portfolios |
| Header Text | Portfolios are a purposeful collection of examinee’s work that shows current proficiency; performance; or knowledge, skills, and abilities (KSAs)—or growth over time. |
| Body Header |  |
| Body Copy | In general, features of portfolios include:   * Printed or electronic copies of text and graphical material that the candidate has produced; other pieces of work; or video and audio recordings of candidate performances, products, and reflections * A broad sampling, as well as a range of types of evidence, skills, and abilities whether current status or growth * Specifications for the criteria and procedures for evaluating the pieces of work |
| Brainshark | Yes |

## 09a More Info

|  |  |
| --- | --- |
| Page Title | What Are Portfolios? |
| Definition | A portfolio assessment can be defined as a purposeful collection of an examinee’s work that is intended to illustrate (a) current proficiency; performance; or knowledge, skills, and abilities (KSAs), or (b) growth over time in proficiency; performance; or KSAs. |
| Characteristic Features | * May be printed or electronic copies of text and graphical material that the candidate has produced; other pieces of work; or video and audio recordings of candidate performances, products, and reflections. * Includes a broad sampling, as well as a range of types of evidence, of the examinee’s proficiency, performance, and KSAs, whether current status or growth. * Criteria and procedures for evaluating the pieces of work are specified for the examinee. |
| Common Contexts | Portfolios are commonly used in educational assessment contexts, including formative and summative, particularly as alternate assessments for students with significant cognitive disabilities. They can also be used in career-related contexts to showcase work products or performances, such as collections of photographs, videos, or other artifacts to demonstrate evidence of proficiency towards a particular job function or skill. |
| Design Variations and Other Considerations | For some portfolios, the examinee selects and assembles the pieces of evidence (e.g., portfolios of artists and architects). For other portfolios, someone else may select and assemble the pieces of evidence (e.g., a teacher who assembles an alternate assessment portfolio for students with significant cognitive disabilities).  For some assessment portfolios, inclusion criteria are highly specified (e.g., for alternate assessment portfolios). For other assessment portfolios, inclusion criteria may be minimal (e.g., portfolios assembled by artists, architects). |
| Response Demands | The examinee (or someone representing the examinee) must select pieces of evidence to include in an assessment portfolio, based on the selection and inclusion specifications. The pieces of evidence are intended to support inferences about current status or growth of proficiency, performance, and KSAs . Examinees may be required to document reflections on the pieces of evidence that provide rationales for inclusion of pieces; explain the thinking and processes behind creating the pieces of evidence; or explain how the pieces of evidence demonstrate either current status or growth of proficiency, performance, and KSAs.    Pieces of evidence that are eligible for inclusion in a portfolio may include written work and other products; audio or video recordings of the examinee that capture a performance, product, demonstration, or reflections; and documentation of external evaluation of the examinee’s performance, product, or demonstration. |
| Evaluation Criteria and Procedures | Written and other concrete forms of evidence are typically scored using one or more rubrics. Performances (whether live or recorded) may be scored using a rubric or evaluated using an observation checklist that identifies specific behaviors or indicators related to the evaluation criteria. Rubrics can be holistic or analytic and evaluation may focus on a narrow set of traits or a more diverse collection of attributes. During evaluation, judges or raters may be able to attach annotations or comment codes to various pieces of evidence as a form of feedback to examinees.    Depending on whether the purpose of the portfolio is to assess an examinee’s current status or growth over time, evaluation criteria may include specific qualities of performances or products (e.g., completeness, consistency) or may instead focus on whether the pieces of evidence demonstrate improvement over time in those qualities. It is especially common when the purpose of the portfolio is to assess growth for the examinee to provide reflection on each piece of evidence that explains how they demonstrate growth or improvement over time. |
| Administration Time | The time required to complete a portfolio varies, depending on the purpose of the portfolio, the range of skills assessed, the number and types of pieces of evidence that are selected, and whether reflection is required. When the purpose of the portfolio is to document growth over time or when the portfolio targets a broad and diverse range of skills, pieces of evidence may be collected over a large span of time (e.g., several months, a year). Conversely, when the purpose of the portfolio is to assess current status or maximal performance or when the portfolio is focused on a relatively narrow range of knowledge and skills, pieces of evidence may be collected within a shorter period of time (e.g, several weeks, a few months). |
| References | Arter, J. A., & Spandel, V. (1992). Using portfolios of student work in instruction and assessment. *Educational Measurement: Issues and Practice*, *11* (1), 36-44.    Koretz, D., Stecher, B., Klein, S., & McCaffrey, D (1994). The Vermont portfolio assessment program: Findings    and implications. *Educational Measurement: Issues and Practice*, *13* (3), 5-16. |

## 09b Exemplars

|  |  |
| --- | --- |
| Page Title in Header | Examples of Portfolios |
| Header Text | Below is a list of innovative examples that demonstrate some of the ways that portfolios may be used as performance assessment. |
| Body Header | [National Board for Professional Teaching Standards (NBPTS) - Adolescent/Young Adult Mathematics Portfolio](http://nbpts.org/for_candidates/the_portfolio?ID=3&x=19&y=8) |
| Body Copy | NBPTS is an independent, nonprofit, nonpartisan and nongovernmental organization formed to advance the quality of teaching and learning.  In this example, Adolescent/Young Adult Mathematics Portfolio, teachers create a portfolio that includes videos, student work, and written commentaries that provide evidence of specific skills and knowledge required by the standards.    National Board for Professional Teaching Standards. Retrieved December 12, 2012 from <http://nbpts.org/> |
| Body Header | [Advanced Placement (AP) - Studio Art: 2-D Design](http://www.collegeboard.com/student/testing/ap/sub_studioart.html?studioart) |
| Body Copy | Through AP's college-level courses and exams, students can earn college credit and advanced placement. While some courses require written exams, Studio Art: 2-D Design requires a portfolio to demonstrate mastery of 2-D design through any two-dimensional medium or process.  The College Board. Retrieved December 12, 2012 from (<http://www.collegeboard.com/student/testing/ap/about.html>) |

# 10 Online Games and Simulated Environments

|  |  |
| --- | --- |
| Page Title in Header | Online Games and Simulated Environments |
| Header Text | Online games and simulated environments are immersive, computer-based settings that feature digital interfaces with interactive, manipulable objects and animations. |
| Body Header |  |
| Body Copy | In general, features of online games and simulated environments include:   * Activities rather than discrete items * Activities set in rich environments that include representations of multiple objects, circumstances, conditions, and interactions, providing context for the problem to be solved * Opportunities to provide immediate, customized formative feedback that can be embedded in the experience * The collection of richer information than is typically collected in traditional assessment contexts, including information about examinee process, work products, and goal attainment |
| Brainshark | Yes |

## 10a More Info

|  |  |
| --- | --- |
| Page Title | What are Online Games and Simulated Environments? |
| Definition | Online games and simulated environments are immersive, computer-based settings that feature digital interfaces with interactive, manipulable objects and animations. These environments often represent detailed fictional or real-world contexts. Educational games and simulations often blur the boundaries between learning and assessment by 1) supporting ubiquitous, unobtrusive, and ongoing assessment opportunities; 2) providing immediate and customized formative feedback; and 3) identifying valuable opportunities for additional learning.  The technology underpinning games and simulated environments allows the collection of rich information about examinee processes, work products, and goal attainment. Such information can support inferences about complex types of knowledge, skills, and attributes, including 21st century competencies like critical thinking, collaboration, and persistence. Multiple learner attributes can be collected and scored to construct multidimensional profiles.  Online games and simulated environments, when used as assessment approaches, are believed to stimulate student interest, engagement, and motivation compared to traditional learning and assessment approaches. |
| Characteristic Features | * Feature digital interfaces with graphic and sometimes audio components and interactive, manipulable objects and animations * Focus on activities rather than discrete items * Typically, offer activities set in rich environments that include representations of multiple objects, circumstances, conditions, and interactions, providing context for the problem to be solved * Can be embedded in daily learning activities rather than isolated as drop-in interruptions to the teaching-learning process * Enable opportunities to provide immediate, customized formative feedback that can be embedded in the experience and provide natural consequences for actions. * May identify additional learning experiences through adaptive activity selection * Are designed to collect richer information than is typically collected in traditional assessment contexts, including information about examinee processes, work products, and goal attainment   As a result of these features, games and simulated environments:   * Are believed to increase student interest, engagement, and motivation * Often blur the boundaries between learning and assessment * Enable ubiquitous, unobtrusive, and ongoing assessment (Behrens et al., 2012); the continuous nature of assessment opportunities is what distinguishes this approach from technology-enhanced items, which are discrete and of finite duration * Can support inferences about more complex types of knowledge, skills, and attributes, including critical thinking, problem solving or decision-making, planning, coordination or collaboration, persistence over time, and interest * Allow development of multidimensional knowledge, skill, or attribute profiles based on a single performance   \*The characteristic features identified are based on currently available research, literature, and examples. This approach to assessment is so new and the field is evolving so rapidly that these features will continue to change. |
| Common Contexts | These are used to simulate real-world phenomena, such as the relationship between fish reproduction and nutrification in a computer model of an aquatic ecosystem; virtual labs, such as virtual dissections of animals; and multiplayer virtual environments, such as online games that support collaborative problem solving. |
| Design Variations and Other Considerations | Games and simulated environments can vary in several different ways in response to design decisions that expand or constrain the problem space, tool space, solution space, and/or response space (Behrens et al., 2012). Problem space refers to the types of problems presented to users, whether open-ended or more constrained. Thus, the problem or goal a user must tackle may be more structured (e.g., carry out an experiment to determine the impact of nitric acid on different types of metal) or less structured (e.g., figure out why fish are dying in a local river).    Similarly, the tools or affordances presented within the environment may be faithful to the types of tools a person would have access to in the “real” world (e.g., surgical tools in a simulation of an operating theatre) or may expand the types of tools to include those that are too expensive or simply not feasible to provide in live environments (e.g., real-time visual display of individual blood cells moving through the patient’s heart).    The solution space, which refers to the set of activities available to the user for solving the problem or achieving the goal, may also be expanded or constrained. For example, the solution space could be constrained to include only limited manipulation of the animation (e.g., varying the “doses” of the independent variable in an experiment) or could be expanded to include a range of possible activities (e.g., designing the experimental setup, interacting with informed agents, or choosing from among multiple settings to explore).    The response space can also be expanded or constrained to produce variation in games and simulations. The response space refers to aspects of user performance within the game or simulation that will be evaluated or scored. For example, user responses may come in the form of selected responses to questions that exist outside the simulation (constrained), but may also include free responses to open-ended writing prompts (expanded). Also, as noted by Behrens et al. (2012), technology allows the consideration of both traditional work products (e.g., users’ written responses), as well as users’ interactions with the system (e.g., actions taken, objects manipulated, agents questioned).    The degree of fidelity or authenticity of the simulation can also be varied, ranging from relatively simple and crude representations to finely textured and detailed representations. However, Behrens et al. (2012) warn against both “over” and “under” simulating aspects of the environment. Thus, simulations should represent material modifications to the problem, tool, solution, or response spaces. However, simulations should not bombard users with unnecessary details that place a burden on cognitive load and introduce construct-irrelevant variance.    Within games and simulations that include a storyline or scenario, the complexity of these settings can vary from simple narratives with one or two characters to complicated settings involving multiple locales, detailed histories, and several characters.    Games and simulations can vary depending on whether they are designed to allow or require collaboration or coordination with other players. Some games are set up as solitary experiences, whereas other games are designed to allow interactions among multiple users (as in multiplayer online games or multiuser virtual environments) or to require collaboration among players in order to evaluate how skillfully users interact with others.    There are still relatively few examples of online games and simulated environments that have been explicitly designed as formal assessments of user knowledge, skills, and attributes. On the other hand, there appear to be a growing number of examples of online games and simulations developed for learning or entertainment purposes. Game developers and simulation researchers have so far focused on trying to use data that is routinely collected within these applications to support inferences about users’ knowledge, skills, and attributes that might inform future learning. However, if games and simulated environments are ever to be used to support high-stakes or summative inferences about what students know and can do, they will have to be designed more purposefully to enable such uses. |
| Response Demands | The response demands of online games and simulations vary according to their purposes and functionality. However, typically users are required to engage in some combination of the following: exploring the environment, identifying goals, inferring rules, manipulating variables, planning, monitoring progress toward goal completion, collaborating or coordinating one’s actions with others, and solving problems or making decisions. |
| Evaluation Criteria and Procedures | Games and simulations provide opportunities to assess knowledge, skills, and attributes (KSAs) that have proven difficult to measure, collect, and score using traditional assessment approaches. Such KSAs include cognitive skills, such as critical or creative thinking; cognitive attributes, such as motivation and persistence; and affective attributes, such as interest or enjoyment. Often, multiple attributes can be evaluated or scored on the basis of a single performance, enabling the construction of multidimensional profiles.    However, the ubiquity of data within digital environments can potentially result in a deluge of information. In this case, the challenge becomes parsing and aggregating the data in ways that can support meaningful inferences about users. For example, the methods for making inferences from log files that record user activity streams are still being developed.    In addition, in order to provide feedback in real time and enable adaptive activity selection, automated scoring of user responses is required. This requirement can, at times, place limitations on the response space as assessment designers must stay within bounds of currently available scoring technology.    Evaluation and scoring can become complicated in the case of multi-user games or simulated environments, particularly when users are either allowed or encouraged to collaborate with others. This is because assessment designers are typically interested in evaluating individual student KSAs, but opportunities for collaboration can obscure individual contributions (Webb, 1995).  Research shows that group-level assessments may not yield scores that are predictive of individual-level ability, even when individual students produce separate work products. In particular, scores from group work tend to over-estimate individual performance and exhibit both ceiling effects and range restriction (Webb, 1993). If this higher performance reflects real student learning, then opportunities for collaboration do not necessarily invalidate multi-user experiences as measures of individual student outcomes. However, if students obtain higher scores from these collaborative experiences simply because their more able counterparts complete most of the work, then viewing scores from multi-user game play as indicators of individual student learning is problematic. |
| Administration Time | Games and simulations are distinguished from other technology-enhanced assessment approaches by their ongoing and continuous nature. Technology allows easy accumulation of data over longer periods of time (e.g., several months or years). Thus, games and simulations can be “administered” to a user on an “anytime” basis, allowing for measurement opportunities that are both distributed (as opposed to massed) and individualized (as opposed to on-demand).    It should also be noted that games and simulations may not be the most efficient way to collect information about basic skills. It would take longer to play through a game to assess basic addition skills than to give a timed test of facts. The game may tell you more about the learner’s ability to apply the skills in context, but if automaticity with facts is the construct of interest, a game or simulation may not be the best assessment approach. |
| References | Behrens, J. T., DiCerbo, K. E., & Ferrara, & S. (2012). Intended and unintended deceptions in the use of     simulations. Paper presented at the Invitational Research Symposium on Technology Enhanced Assessment,  Washington, DC.  Chung, G. K. W. K. & Baker, E. L. (1997). *Year 1 technology studies: implications for technology in assessment* (CSE Technical Report No. 459). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing.  Quellmalz, E. S., Timms, M.J., & Schneider, S. A. (2009). *Assessment of student learning in science simulations and games*. Washington, DC: National Research Council.  Rupp, A. A., Levy, R., DiCerbo, K. E., Sweet, S. J., Crawford, A. V., Calico, T., Benson, M., Fay, D., Kunze, K. L., Mislevy, R. J., Behrens, J. T. (2012). Putting ECD into practice: The interplay of theory and data in evidence models within a digital learning environment. *Journal of Educational Data Mining*, 4(1), 49-110.  Rupp, A. A., Nugent, R., & Nelson, B. (2012). Evidence-centered design for diagnostic assessment within digital learning environments: Integrating modern psychometrics and educational data mining. *Journal of Educational Data Mining*, 4(1), 1-10.  Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. In S. Tobias & J. D. Fletcher (Eds.), *Computer games and instruction* (pp. 503-524). Charlotte, NC: Information Age Publishers. |

## 10b Exemplars

|  |  |
| --- | --- |
| Page Title in Header | Examples of Online Games and Simulated Environments |
| Header Text | Below is a list of innovative examples that demonstrate some of the ways that online games and simulated environments may be used as performance assessment. |
| Body Header | [**Quest Atlantis: Taiga Park** **(Atlantis Remixed)**](http://atlantisremixed.org/) |
| Body Copy | Atlantis Remixed (ARX) is an international learning and teaching project that uses a 3D multi-user environment to immerse children, ages 9-16, in educational tasks. ARX combines strategies used in commercial games with lessons from educational research on learning and motivation.  Retrieved December 11, 2012 from [http://atlantisremixed.org](http://atlantisremixed.org/) |
| Body Header | [**River City**](http://muve.gse.harvard.edu/rivercityproject/view/rc_views_interface.htm) **(Harvard University, Arizona State University)** |
| Body Copy | As visitors to River City, students travel back in time, bringing their 21st century skills and technology to address 19th century problems. Based on authentic historical, sociological, and geographical conditions, River City is a town besieged with health problems. Students work together in small research teams to help the town understand why residents are becoming ill. Students use technology to keep track of clues that hint at causes of illnesses, form hypotheses, develop controlled experiments to test their hypotheses, and make recommendations based on the data they collect, all in an online environment.    Retrieved December 11, 2012 from <http://muve.gse.harvard.edu/rivercityproject/view/rc_views_interface.htm> |
| Body Header | [**Hurricane Landfall**](http://www.dd.ucar.edu/)**(National Center for Atmospheric Research)** |
| Body Copy | The Hurricane Landfall game teaches players about interactions between natural hazards and human decisions in a Gulf Coast barrier island community. It is a strategy and negotiation computer game intended for use in undergraduate classes and is designed for four players who are connected to one another via the Internet.      Retrieved December 11, 2012 from <http://www.dd.ucar.edu/> |
| Body Header | [**Save Patch**](http://archive.org/details/SavePatch)**(CRESST, USC’s Game Innovation Lab)** |
| Body Copy | This middle-school game draws on topics related to rational numbers, including identification of unit size, the numerator, denominator, and addition of fractions. The objective of the game is to help the game character (Patch) jump over obstacles (e.g., spikes, lava, quicksand) and move from block to block to reach the last “X” block (the final goal). To do this, players need to compute the distance of the jump, place trampolines on the blocks, and add enough coils to the trampolines to make Patch bounce.    Retrieved December 11, 2012 from <http://archive.org/details/SavePatch> |