Beyond Multiple-Choice with Digital Assessments

SPECIAL ISSUE

Advancing Beyond Multiple Choice eAssessment

by Okan Bulut

During the past decade, K-12 education systems have been increasingly relying on digital forms of educational assessments. Therefore, digital assessments have been integrated into the instructional process and curriculum in multiple ways to promote student learning inside and outside the classroom. When developing digital assessments, one of the most important elements is the type of items used in the assessment. As technological innovations continue to change the type of tasks we can measure using digital assessments, new types of items also emerge. Items in digital assessments can go beyond the limits of what can be measured on a paper-and-pencil assessment with traditional multiple-choice items. This article will provide a summary of different item types in the context of digital assessments and discuss how they differ from traditional item types.

Beyond Multiple-Choice with Digital Assessments

SPECIAL ISSUE

Advancing Beyond Multiple Choice eAssessment

by Okan Bulut

During the past decade, K-12 education systems have been increasingly relying on digital forms of educational assessments—such as summative and formative assessments, screening tests, and progress monitoring tools—as a means for measuring student learning. Teachers often use digital assessments as a summative assessment tool to evaluate and grade students’ overall understanding of the content covered in the lessons. Also, teachers use digital assessments as a formative assessment tool to monitor students’ learning progress throughout the school year, provide students with timely feedback on their performance [1, 2], and adjust their instruction to address individual student needs [3]. Similarly, e-tutoring programs designed for K-12 students include online assessments that allow students to self-assess their learning. Overall, digital assessments have been integrated with the instructional process and curriculum in multiple ways to promote student learning inside and outside the classroom.

When developing digital assessments, one of the most important elements is the type of items used in the assessment. Unlike paper-and-pencil assessments that require teachers to manually score the items, items in digital assessments (both selected-response and constructed-response items) can be scored automatically. With increased efficiency in scoring, teachers can evaluate student achievement more frequently and monitor students’ learning progress more closely. Furthermore, items used in digital assessments can bring additional value to the assessment process through log files and process data of students’ actions. For example, interactive items in a digital assessment produce process data (clicks and sequence of actions taken by students) that can be very useful in understanding students’ test-taking and problem-solving behaviors. This article will provide a summary of different item types in the context of digital assessments and discuss how they differ from traditional item types.

Multiple-Choice Testing

Digital assessments often include selected-response items (for example, true/false, multiple-choice, and matching) where students answer the items by choosing a response option provided by the teacher or test developer. Among all types of selected-response items, multiple-choice items are undeniably the most common format. Despite the ongoing controversies around standardized testing, multiple-choice testing remains one of the most effective and enduring forms of educational assessment [4]. Multiple-choice items are widely used in both paper-and-pencil and digital assessments because: (a) they are easy to score; (b) there is no room for subjectivity in scoring; (c) the neatness of handwriting, penmanship, and writing skills do not influence the scoring process; and (d) teachers can assess more content in a shorter amount of time using several multiple-choice items.

Although multiple-choice items can be a reliable and efficient way to evaluate large numbers of students, their utility in education remains subject to criticism due to several constraints and difficulties in practice. First, developing multiple-choice items with clear stems (that is, question statement) and plausible distractors (that is, incorrect response options) is a highly time-consuming and challenging task [4]. Second, depending on the quality of distractors, multiple-choice items can be very easy for students to guess the right answer [5]. Third, multiple-choice items often focus on low-level knowledge and skills (for example, remembering factual knowledge) and thus may not be well suited for assessing complex skills such as critical thinking, reasoning, and problem solving [6]. Therefore, most teachers would concur that alternative item formats, such as constructed-response items, should also be included in educational assessments to measure a wider range of knowledge, skills, and competencies [7].

Constructed-Response Items

Constructed-response items (for example, fill-in-the-blank items, short-answer items, and essays) require students to develop their answers by using their knowledge, reasoning skills, and critical thinking abilities [8]. That is, students construct their own responses without the benefit of any response options. Constructed-response items may be simple (comparing and contrasting two types of habitats for a particular type of plant) or complex (writing a short essay about the potential effects of climate change) in nature. Compared with multiple-choice items, constructed-response items are more effective assessment tools because not only do these items encourage students to think deeply about their answers they also reduce the possibility of guessing since students are not given any response options to select from.

Unlike multiple-choice items that can be easily scored using an answer key, constructed-response items are harder to score because they rely on human raters [7]. Teachers are responsible for manually scoring students' responses to constructed-response items in classroom assessments, while external graders manually score students' responses to high-stakes assessments (for example, advanced placement exams in the United States). Raters typically use either an analytic rubric or a holistic rubric to decide how many points to award to the response [9]. An analytic scoring rubric is a detailed scoring guide that includes several evaluation categories and specifies the number of points to award for each evaluation category. Compared with holistic scoring, analytic scoring tends to yield more consistent scores from one rater to another [9]. Figure 1 shows an example of an analytic rubric designed for scoring student essays. Unlike an analytic rubric, a holistic rubric usually defines a typical student response at each score level, without creating multiple evaluation categories. Figure 2 shows an example of a holistic rubric for scoring reading activities in grades 5–8.

Figure 1. An analytic rubric for scoring essays (<http://www.readwritethink.org/>).

Traits	4	3	2	1
Focus & Details	There is one clear, well-focused topic. Main ideas are clear and are well supported by detailed and accurate information.	There is one clear, well-focused topic. Main ideas are clear but are not well supported by detailed information.	There is one topic. Main ideas are somewhat clear.	The topic and main ideas are not clear.
Organization	The introduction is inviting, states the main topic, and provides an overview of the paper. Information is relevant and presented in a logical order. The conclusion is strong.	The introduction states the main topic and provides an overview of the paper. A conclusion is included.	The introduction states the main topic. A conclusion is included.	There is no clear introduction, structure, or conclusion.
Voice	The author's purpose of writing is very clear, and there is strong evidence of attention to audience. The author's extensive knowledge and/or experience with the topic is/are evident.	The author's purpose of writing is somewhat clear, and there is some evidence of attention to audience. The author's knowledge and/or experience with the topic is/are evident.	The author's purpose of writing is somewhat clear, and there is evidence of attention to audience. The author's knowledge and/or experience with the topic is/are limited.	The author's purpose of writing is unclear.

Figure 2. A holistic rubric for scoring reading activities in grades 5–8
[\(https://exemplars.com/\)](https://exemplars.com/).

Reading Reading multiple sources or texts Inference Interpretation: i.e., literal, implied, figurative language Evaluates information and ideas	
Novice	<ul style="list-style-type: none"> • Reads information from several texts, only if supported by directions/activities. • Understands basic information contained in texts, with assistance. • Uses models for evaluation or questioning texts, with assistance if necessary.
Apprentice	<ul style="list-style-type: none"> • Gathers details or additional information by reading several texts. • Develops a comprehensive, yet literal interpretation of materials collected. • Evaluates or questions information with assistance
Practitioner	<ul style="list-style-type: none"> • Develops an extended understanding of the topic based on using multiple sources (primary and secondary texts, literature, art, artifacts). • Infers meanings and interprets age appropriate texts. • Understands both literal and figurative meaning in a text. • Begins to evaluate or question sources independently or in a peer group.
Expert	<ul style="list-style-type: none"> • Synthesizes a varied body of information including primary and secondary sources, literature, artifacts, etc. • Interprets and uses challenging texts. Comprehends mature vocabulary. • Understands and integrates the literal and figurative meanings of a text; reads textual materials with empathy. • Analyzes sources critically and objectively, cross-checking for differences, conflicts, etc.

A major challenge for educational assessments including constructed-response items is the duration and cost of the scoring process. Typically, two or more highly trained raters spend large amounts of time to score constructed-response items objectively and consistently. If double-blind marking is used, two raters independently score the items and then compare the results. If the raters substantially disagree on the score for a given answer, a third rater may be invited to review the same answer. Alternatively, each rater can score a random sample of items, instead of using double-blind marking. Although raters receive training on how to score constructed-response items using a scoring rubric, it may not be possible to eliminate errors in the human scoring of constructed-response items. For example, one potential source of bias referred to as the “halo effect” may occur if extraneous factors (such as penmanship, prior experience with a student) influence the scoring. Another example would be the inconsistent scoring of long essays due to raters’ fatigue during the scoring process.

Automated Essay Scoring

One promising alternative to the human scoring of constructed-response items is to implement automated scoring systems powered by artificial intelligence (AI). To score students’ written responses automatically, digital assessments can employ cutting-edge techniques in Natural Language Processing (NLP)—a branch of AI that deals with the computational modeling of human language using computer technology. The primary goal

of NLP is to transform human input (for example, text or voice) into a form that computers can understand. In the context of scoring constructed-response items, NLP can be used to create an automated essay scoring (AES) system—a computer program that can automatically score responses given to different types of items ranging from short-answer items to long essays. A well-designed AES system can score thousands of student responses quickly, with a high degree of precision. Furthermore, an AES system can produce actionable feedback on how students could improve their responses [10].

There are several steps involved in building an AES system. First, the system requires a large training data set including responses to constructed-response items that have been pre-scored by human raters. Second, the system finds and extracts linguistic features from students' responses. Generally, human raters use a rubric to score constructed-response items based on the relevance of response to the item, the organization of the response, and lower-level errors (for example, grammar, typos, and punctuation) [10]. Unlike human raters, AES does not try to “understand” the content of the response. Instead, it looks for consistent linguistic patterns across the responses. Third, the system builds a scoring model by matching the scores assigned by human raters and the linguistic patterns identified in the previous step. Fourth, the AES system applies the scoring model to automatically score the responses of a new group of students [7].

There are several benefits of scoring constructed-response items with an AES system, such as faster scoring of written responses and lower costs for the scoring process [7]. More importantly, AES improves the consistency of scoring because external factors (such as halo effect and rater fatigue) cannot influence the scoring process. It must be noted that the success of an AES system depends on several conditions. For example, the training data set that the AES system is built upon must be large enough in terms of the number of students. The larger the data set, the more accurately the AES system can capture differences in student responses. Furthermore, scores assigned by the human raters in the training data set must be highly consistent. If the human raters score the items inconsistently (that is, high scoring error), the AES system cannot understand the scoring mechanism and thus yields inconsistent scores.

Over the past decade, there has been a growing interest in the development of advanced AES systems for scoring student responses. For example, The William and Flora Hewlett Foundation sponsored two public contests, called the Automated Student Assessment Prize, in 2012 [11]. Competitors from different testing organizations were tasked with developing an effective AES system that could reproduce the scores assigned by human raters for 22,029 student essays. The essays were written by students in grades 7, 8, and 10 in the United States. The results of these contests indicated that the AES systems yielded highly accurate scores that were similar to those produced by a group of highly trained

human raters. Similar results from the following AES-focused events have led to useful technical advances and applications for the task of automated scoring. Today, there is sufficient empirical evidence to justify the use of AES for scoring constructed-response items in both low-stakes and high-stakes assessments in education. Many testing organizations use AES to score students' responses to constructed-response items, such as American Institutes for Research, Educational Testing Service, ACTNext, and the Australian Council for Educational Research.

Technology-Enhanced Items

In addition to traditional item formats (for example, multiple-choice and constructed-response items), there are also more innovative item formats that harness the power of computer technology, which are often referred to as “technology-enhanced items.” These items include a wide range of types, such as items including interactive simulations, items with sound and video animations, and multistep items with integrated tasks or scenarios. Figure 3 shows an example of an interactive science item. Using this item, students can investigate the pH of everyday liquids (milk, chicken soup, and coffee). The interactive elements of the item (the water valve) enable students to test how adding more of a liquid or diluting with water would affect pH.

More advanced examples of technology-enhanced items include digital simulations in which students view a virtual environment (a science lab) and complete a set of tasks (running a series of science experiments). These items can be presented to students within an online test or immersive learning environments through virtual reality (VR) headsets [12]. Furthermore, technology-enhanced items with video and audio recording capabilities can help educators collect data beyond traditional multiple-choice and constructed-response items. Therefore, digital assessments using such interactive item types may encourage students to stay engaged with the assessment environment and thereby improve the assessment of their knowledge and skills.

Figure 3. An example of an interactive simulation item
(<https://phet.colorado.edu/en/simulation/ph-scale-basics>).

Other benefits of using technology-enhanced items in digital assessments include improved measurement quality with more authentic tasks [13], the opportunity to measure higher-level cognitive abilities [14], and the possibility of evaluating the response process in addition to the response outcome [15]. Furthermore, technology-enhanced items provide the teacher or assessment specialist with the opportunity to observe a wide range of student behaviors beyond a correct or incorrect response [16]. For example, the number of clicks on the screen, the number and sequence of actions taken, and the time spent on the item can inform both teachers and assessment specialists about students' level of engagement with the digital assessment.

Conclusion

With the integration of advanced technologies into teaching and learning, the use of digital assessments will continue to increase in school systems around the world. Both students and teachers are likely to benefit from the increased use of digital assessments in education. Students may appreciate the opportunity to demonstrate their learning through more novel and engaging items, instead of dealing with traditional item types such as multiple-choice items. Through digital assessments, teachers may spend less time on grading items and focus more on the use of assessment results to adapt their teaching strategies to student needs. The future of digital assessments is likely to include more

novel and innovative item types that can be further personalized to each student's learning needs. The increased need for personalization in both learning and assessment is likely to remain a powerful catalyst for the development of more innovative items in digital assessments.

References

- [1] Bulut, O. et al. [Effects of digital score reporting and feedback on students' learning in higher education](#). *Frontiers in Education* 4, 65 (2019), 1–16.
- [2] Bulut, O. et al. [Guidelines for generating effective feedback from e-assessments](#). *Hacettepe University Journal of Education* 35, Special Issue (2020), 60–72.
- [3] Bulut, O. et al. [An intelligent recommender system for personalized test administration scheduling with computerized formative assessments](#). *Frontiers in Education* 5 (2020), 1–11.
- [4] Gierl, M. J., Bulut, O., et al. [Developing, analyzing, and using distractors for multiple-choice tests in education: A comprehensive review](#). *Review of Educational Research* 87, 6 (2017), 1082–1116.
- [5] Novacek, P. [Confidence-based assessments within an adult learning environment](#). International Association for Development of the Information Society. 2013.
- [6] Shute, V. J., and Rahimi, S. [Review of computer-based assessment for learning in elementary and secondary education](#). *Journal of Computer Assisted Learning* 33, 1 (2017), 1–19.
- [7] Gierl, M. J. et al. [Automated essay scoring and the future of educational assessment in medical education](#). *Medical Education* 48, 10 (2014), 950–962
- [8] Tankersley, K. *Tests That Teach: Using Standardized Tests to Improve Instruction*. ASCD, 2007.
- [9] Livingston, S. A. [Constructed-response test questions: Why we use them; how we score them](#). *R&D Connections* 11 (2009). Educational Testing Service.
- [10] Madnani, N. and Cahill, A. [Automated scoring: Beyond natural language processing](#). In *Proceedings of the 27th International Conference on Computational Linguistics* 2018, 1099–1109.
- [11] [The William and Flora Hewlett Foundation: Automated Essay Scoring](#). 2012.
- [12] Wools, S. et al. The validity of technology enhanced assessments—threats and opportunities. In *Theoretical and Practical Advances in Computer-Based Educational Measurement*. Springer, Cham, Switzerland, 2019, 3–19.
- [13] Parshall, C. G. and Harmes, J. C. Improving the quality of innovative item types: Four tasks for design and development. *Journal of Applied Testing Technology* 10, 1 (2009), 1–20.
- [14] Wendt, A. et al. Assessing critical thinking using a talk-aloud protocol. *CLEAR Exam Review* 18, 1 (2007), 18–27.

[15] Behrens, J. T., DiCerbo, K. E. Technological implications for assessment ecosystems: Opportunities for digital technology to advance assessment. Paper written for the Gordon Commission on the Future of Assessment in Education. 2012.

[16] Rupp, A. A., Levy, R., et al. 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 (2012), 49–110.

About the Author

Okan Bulut is an associate professor in the Measurement, Evaluation, and Data Science program and a researcher at the Centre for Research in Applied Measurement and Evaluation (CRAME) at the University of Alberta. Dr. Bulut teaches courses on psychometrics, educational measurement, and statistical modeling. Also, he gives workshops and seminars on advanced topics, such as data mining, big data modeling, data visualization, and statistical data analysis using software programs like R and SAS. His current research interests include educational data mining, big data modeling, computerized/digital assessments, psychoeducational assessments, and statistical programming using the R programming language.

DOI: 10.1145/3472394