



Technology and Testing

Edys S. Quellmalz, *et al.*
Science **323**, 75 (2009);
DOI: 10.1126/science.1168046

The following resources related to this article are available online at www.sciencemag.org (this information is current as of January 11, 2009):

Updated information and services, including high-resolution figures, can be found in the online version of this article at:

<http://www.sciencemag.org/cgi/content/full/323/5910/75>

A list of selected additional articles on the Science Web sites **related to this article** can be found at:

<http://www.sciencemag.org/cgi/content/full/323/5910/75#related-content>

This article appears in the following **subject collections**:

Education

<http://www.sciencemag.org/cgi/collection/education>

Information about obtaining **reprints** of this article or about obtaining **permission to reproduce this article** in whole or in part can be found at:

<http://www.sciencemag.org/about/permissions.dtl>



judged by the transfer of what is learned in instruction to what is done in practical operations. Quantitative attempts to deal with this issue have used transfer effectiveness ratios to balance the cost of simulator time to the cost of using real equipment (34). "Isoperformance" curves have also been developed to help instruction designers identify points at which different combinations of training inputs produce equivalent performance output with minimal costs (35).

Conclusion

Military organizations have their own perspectives and emphases, but the techniques and technologies that they have developed in the following areas, among others, continue to be of interest and value beyond the military.

Training technology. After reviewing the issue of tailoring instruction to the needs of each learner, Scriven (36) concluded that it was both an educational imperative and an economic impossibility. Continued DOD interest in developing CAI arises from an expectation that computer technology will make this imperative affordable (11). The results from the 1960s on have been instructional technologies that adjust the pace, sequence, and difficulty of tasks so that learning is accelerated, allowing learners to focus on what they need to learn rather than what they already know.

Instructional efficiency. Military organizations, which assume responsibility for individuals from enlistment through retirement, have concentrated on the development of techniques and principles that increase instructional efficiency and assess the cost-effectiveness of alternate approaches.

Collective performance. Instructional technology for crews, teams, and units is a particular concern of military organizations. Techniques for developing shared mental models, conducting group assessments, encouraging collaboration, and measuring the competence, productivity, and readiness of collectives should be of value to all sectors.

Research and development. The military continues to invest substantially in research and development for instructional technology. Some of its instructional technology programs, particularly those in skill-training areas, have been transferred to specific civilian applications. However, its open nonproprietary development of techniques, technologies, and capabilities in nonclassified areas, particularly those of CAI and simulation, has influenced instructional practice in all sectors.

References and Notes

- J. Kiszely, "Post-Modern Challenges for Modern Warriors" (Shrivenham Paper No. 5, Defence Academy of the United Kingdom, Swinton, Wiltshire, UK, 2007).
- J. E. Coulson, Ed., *Programmed Learning and Computer-Based Instruction* (Wiley, New York, 1962).
- E. H. Galanter, Ed., *Automatic Teaching: The State of the Art* (Wiley, New York, 1959).
- J. D. Fletcher, M. R. Rockway, in *Military Contributions to Instructional Technology*, J. A. Ellis, Ed. (Praeger, New York, 1986), pp. 171–222.
- Office of Technology Assessment, U.S. Congress, *Power On! New Tools for Teaching and Learning* (OTA-SET-379, Government Printing Office, Washington, DC, 1988), p. 158.
- D. Bitzer, P. Braunfield, W. Lichtenberger, *IEEE Trans. Educ.* **4**, 157 (1961).
- R. C. Atkinson, H. A. Wilson, Eds., *Computer-Assisted Instruction: A Book of Readings* (Academic, New York, 1969).
- A. S. Gibbons, P. G. Fairweather, in *Training and Retraining: A Handbook for Business, Industry, Government, and the Military*, S. Tobias, J. D. Fletcher, Eds. (Macmillan Reference, New York, 2000), pp. 410–442.
- J. D. Ford, D. A. Slough, R. E. Hurlock, *Computer Assisted Instruction in Navy Technical Training Using a Small Dedicated Computer System: Final Report* (Research Rep. No. SRR 73-13, Navy Personnel Research and Development Center, San Diego, CA, 1972).
- J. F. Vinsonhaler, R. K. Bass, *Educ. Technol.* **12**, 29 (1972).
- J. D. Fletcher, *Int. J. Cogn. Ergon.* **5**, 317 (2001).
- W. R. Uttal, in *Programmed Learning and Computer-Based Instruction*, J. E. Coulson, Ed. (Wiley, New York, 1962), pp. 171–190.
- J. R. Carbonell, *IEEE Trans. Man Mach. Syst.* **11**, 190 (1970).
- J. Psotha, L. D. Massey, S. A. Mutter, Eds., *Intelligent Tutoring Systems: Lessons Learned* (Lawrence Erlbaum Associates, Hillsdale, NJ, 1988).
- R. Luckin, K. R. Koedinger, J. Greer, Eds., *Artificial Intelligence in Education* (IOS Press, Amsterdam, 2007).
- P. Dodds, J. D. Fletcher, *J. Educ. Multimed. Hypermedia* **13**, 391 (2004).
- S. Tobias, T. D. Duffy, Eds., *Constructivist Theory Applied to Education: Success or Failure?* (Taylor and Francis, New York, 2008).
- W. H. Wulfek, S. K. Wetzel-Smith, E. Baker, in *Assessment of Problem Solving Using Simulations*, J. Dickieson, W. Wulfek, H. F. O'Neil, Eds. (Taylor and Francis—Lawrence Erlbaum Associates, Florence, KY, 2007), pp. 223–238.
- in *Performance Enhancement in High Risk Environments*, J. V. Cohn, P. E. O'Connor, Eds. (Praeger, in press).
- L. W. Anderson, D. R. Krathwohl, Eds., *A Taxonomy for Learning, Teaching, and Assessing: A Taxonomy of Educational Objectives* (Allyn and Bacon, Columbus, OH, 2001).
- A. Lesgold, S. Lajoie, M. Bunzo, G. Eggan, *SHERLOCK: A Coached Practice Environment for an Electronics Troubleshooting Job* (Learning Research and Development Center, University of Pittsburgh, Pittsburgh, PA, 1988).
- G. Klein, in *Aircrew Training and Assessment: Methods, Technologies, and Assessments*, H. F. O'Neil, D. Andrews, Eds. (Lawrence Erlbaum Associates, Mahwah, NJ, 2000), pp. 165–195.
- A. C. Williams, in *Aviation Psychology*, S. N. Roscoe, Ed. (Iowa State Univ. Press, Ames, IA, 1980), pp. 11–30.
- C. J. Biddle, *Fighting Airman: The Way of the Eagle* (Doubleday, Garden City, NY, 1968).
- H. F. O'Neil, D. Andrews, Eds., *Aircrew Training and Assessment: Methods, Technologies, and Assessments* (Lawrence Erlbaum Associates, Mahwah, NJ, 2000).
- W. Bennett Jr., C. E. Lance, D. J. Woehr, *Performance Measurement: Current Perspectives and Future Challenges* (Routledge, New York, 2006).
- J. A. Cannon-Bowers, R. Oser, D. L. Flanagan, in *Teams: Their Training and Performance*, R. W. Swezey, E. Salas, Eds. (Ablex, Norwood, NJ, 1992), p. 355.
- B. M. Huey, C. D. Wickens, *Workload Transition: Implications for Individual and Team Performance* (National Academies Press, Washington, DC, 1993).
- P. F. Gorman, in *Proceedings of the 1991 Summer Computer Simulation Conference*, D. Pace, Ed. (Society for Computer Simulation, Baltimore, MD, 1991), pp. 1181–1186.
- J. E. Morrison, L. L. Meliza, *Foundations of the After Action Review Process* (U.S. Army Research Institute for the Behavioral and Social Sciences, Alexandria, VA, 1999).
- E. A. Alluisi, *Hum. Factors* **33**, 343 (1991).
- J. A. Thorpe, in *Proceedings of the Ninth InterService/Industry Training Systems Conference* (American Defense Preparedness Association, Arlington, VA, 1987), pp. 492–501.
- D. H. Andrews, L. A. Carroll, H. H. Bell, *The Future of Selective Fidelity in Training Devices* (AL/HR-TR-1995-0195, Armstrong Laboratory, Aircrew Training Research Division, Williams Air Force Base, AZ, 1996).
- S. N. Roscoe, B. H. Williges, in *Aviation Psychology*, S. N. Roscoe, Ed. (Iowa State University Press, Ames, IA, 1980), pp. 182–193.
- M. B. Jones, R. S. Kennedy, *Hum. Factors* **38**, 167 (1996).
- M. Scriven, in *Systems of Individualized Education*, H. Talmage, Ed. (McCutchan, Berkeley, CA, 1975), pp. 199–210.

37. Conclusions and opinions expressed are those of the author and do not represent official positions of DOD.

10.1126/science.1167778

PERSPECTIVE

Technology and Testing

Edys S. Quellmalz^{1*} and James W. Pellegrino²

Large-scale testing of educational outcomes benefits already from technological applications that address logistics such as development, administration, and scoring of tests, as well as reporting of results. Innovative applications of technology also provide rich, authentic tasks that challenge the sorts of integrated knowledge, critical thinking, and problem solving seldom well addressed in paper-based tests. Such tasks can be used on both large-scale and classroom-based assessments. Balanced assessment systems can be developed that integrate curriculum-embedded, benchmark, and summative assessments across classroom, district, state, national, and international levels. We discuss here the potential of technology to launch a new era of integrated, learning-centered assessment systems.

A new generation of technology-enabled assessments offers the potential for transforming what, how, when, where, and why testing occurs. Powered by the ever-increasing capabilities of technology, these 21st-century ap-

proaches to assessment expand the potential for tests to both probe and promote a broad spectrum of human learning, including the types of knowledge and competence advocated in various recent policy reports on education and the

Education & Technology

economy [e.g., (1, 2)]. The use of assessment to support the attainment of such goals will require interdisciplinary partnerships and considerable additional research and development. It will also demand major shifts in educational policies regarding the use of assessment data for various purposes, including student, teacher, and system-level accountability. Here, we look at both the current state of technology use in testing and some of the emergent cases that have used technology to push the envelope with regard to educational assessment.

Technology Applications in Current Large-Scale Assessment Programs

In large-scale assessment programs such as those run by states or nations and by major testing companies, technology currently supports a myriad of assessment functions, including test development, delivery, adaptation, scoring, and reporting [e.g., (3, 4)]. Authoring shells that guide the process of item writing and item banks aligned to content standards enable efficient development and assembly of items into comparable test forms. Online administration eliminates costs for shipping, tracking, and collecting print booklets while simultaneously introducing other logistical complexities related to equipment and security. Computer scoring provides rapid return of results and generation of reports tailored to multiple audiences. Flexible administration times and locales shift annual, on-demand testing to interim and just-in-time challenges.

Online testing now occurs in numerous international, national, and state assessment programs. The 2009 Programme for International Student Assessment (PISA) will include electronic texts to test reading, and in 2006 PISA conducted a pilot of computer-based assessment in science (5). The National Assessment of Educational Progress (NAEP) studied online versions of mathematics and writing tests in preparation for transitioning NAEP to electronic administrations in the near future (6). Currently, more than 27 states have operational or pilot versions of online tests for their statewide or end-of-course exams. This includes Oregon, which pioneered online statewide assessment, North Carolina, Utah, Idaho, Kansas, Wyoming, and Maryland. The landscape is changing rapidly, as is the growth of computer-administered tests. For example, the Educational Testing Service (ETS) estimates that more than 4 million people will take ETS-developed tests on computer in 2008. Those tests range from the Graduate Record Exam (GRE), to state and national teacher competency tests, to selected areas of the Advanced Place-

ment Program. This is representative of what is happening industry-wide and on a much larger scale.

Computerized adaptive testing (CAT) procedures, in which items are selected based on the examinee's prior response history and an underlying measurement model of proficiency, have been developed to reduce testing time and examinee burden. However, their use in high-stakes testing contexts has been confined largely to admission, professional certification, and credentialing exams. Despite requests to use CAT in state testing programs, regulations of the No Child Left Behind (NCLB) legislation have prohibited their implementation for assessment of student academic achievement in reading, mathematics, and science.

A transformative advance in large-scale testing programs is the machine scoring of essays and constructed responses, including testing programs for the military, industry training, higher education admissions, and statewide kindergarten through grade 12 achievement testing. Computerized scoring of free responses uses complex statistical methods and techniques such as latent semantic analysis (LSA) (7). Test publisher Pearson is in its second year of using Knowledge Analysis Technologies, based on LSA techniques, to pilot the automated scoring of 46,000 brief constructed responses for the Maryland School Assessment science test. ETS has developed E-rater for scoring essays and C-rater for scoring constructed responses and has deployed them in a variety of high-stakes testing programs such as the GRE.

Klein (8) recently reviewed the literature on automated scoring methods and presented results from a study comparing hand and machine scoring of college-level, open-ended items of the type found on the Collegiate Learning Assessment. Findings across studies using a variety of machine scoring methods consistently show comparability of human and machine scoring (~0.85 score correspondence) at levels approximating the agreement between two human scorers (~0.86) and sufficient to warrant using computerized scoring alone, or as an augmentation to human scoring.

In summary, the current genre of online testing applications remains focused on (i) electronic delivery of conventional selected-response and constructed-response item formats, (ii) automation of existing processes, and (iii) comparability of scores and interpretation of results across computer-based and paper forms. A next generation of assessments, however, is attempting to move beyond this limited framing of assessment issues and overcome many of the limitations of conventional testing practices. A goal is to harness technology to enable assessment of those aspects of cognition and performance that are complex and dynamic and that were previously impossible to assess directly. Such work involves reconceptualizing assessment design and use and tying assessment more directly to the processes and contexts of learning and instruction.

Toward the Next Generation of Technology-Enabled Assessment

Across the disciplines, technologies have expanded the phenomena that can be investigated,

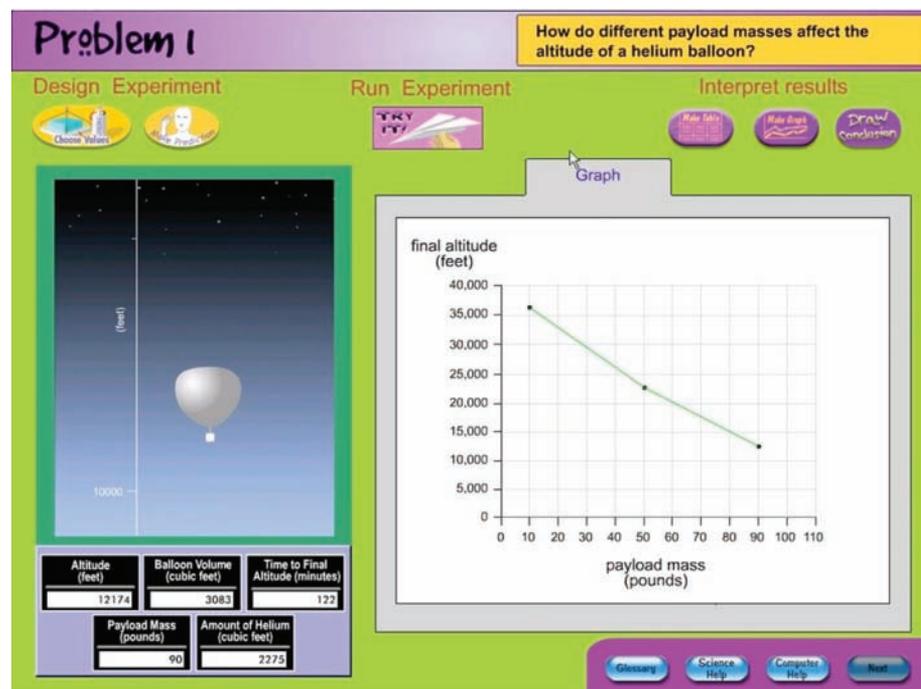


Fig. 1. The ETS simulation model (10).

¹WestEd, 400 Seaport Court, Suite 222, Redwood City, CA 94063, USA. ²Learning Sciences Research Institute, MC 057, University of Illinois at Chicago, 1007 West Harrison Street, Chicago, IL 60607-7137, USA.

*To whom correspondence should be addressed. E-mail: equellm@wested.org



the nature of argumentation, and the use of evidence. They allow representation of domains, systems, models, and data, and their manipulation, in ways that previously were not possible. Dynamic models of ecosystems or molecular structures help scientists visualize and communicate complex interactions. Models of population density permit investigations of economic and social issues. This move from static to dynamic models has changed the nature of inquiry among professionals and the way that academic disciplines can be taught and tested. Moreover, the computer's ability to capture student inputs permits collecting evidence of processes such as problem-solving sequences and strategy use as reflected by information selected, numbers of attempts, and time allocation. Such data can be combined with statistical and measurement algorithms for the extraction of patterns associated with varying levels of expertise [e.g., (9)]. In addition, technology can be used to design new forms of adaptive testing that integrate diagnosis of errors with student and teacher feedback.

Large-scale assessment programs. Information and communications technologies such as web browsers, word processors, editing, drawing, and multimedia programs support research, design, composition, and communication processes. These same tools can expand the cognitive skills that can be assessed, including planning, drafting, composing, and revision. For example, the NAEP writing assessment in 2011 will require the use of word processing and editing tools to compose essays. In professional testing, architecture examinees use computer-assisted design programs as part of their licensure assessment. The challenge offered by

such technology-based presentation and data-capture contexts now lies in the analysis of complex forms of data and their meaningful interpretation relative to models of the underlying components of competence and expertise.

Science assessment is perhaps leading the way in exploring the presentation and interpretation of complex, multifaceted problem types and assessment approaches. In 2006, PISA pilot-tested a Computer-Based Assessment of Science specifically to test knowledge and inquiry processes not assessed in the paper-based booklets. Their assessment included student exploration of the genetic breeding of plants. At the state level, Minnesota has an online science test with tasks engaging students in simulated laboratory experiments or investigations of phenomena such as weather or the solar system. ETS pioneered the design of technology-based assessments for complex learning and performance (10). An example of the type of item that Bennett *et al.* evaluated in (10) is shown in Fig. 1.

Students were presented with a scenario involving a helium balloon and asked to determine how different payload masses affect the altitude of the balloon. They could design an experiment, manipulate parameters, run their experiment, record their data, and graph the results. Figure 1 also shows the types of data that might be obtained by a student and plotted before reaching a conclusion and writing a final response. The 2009 NAEP Science Framework and specifications drew upon ETS work and other research in developing their rationale for the design and pilot testing of Interactive Computer Tasks to test students' ability to engage in inquiry practices. Such innovative items will be included in the upcoming 2009 NAEP science administration.

Large-scale testing programs such as those mentioned above are just beginning to explore the possibilities of using dynamic, interactive tasks for obtaining evidence of student content knowledge and reasoning. However, in the realm of high-stakes assessment for NCLB accountability, a number of regulatory, economic, and logistical issues still constrain the breadth and depth of the content and performance standards that are assessed in annual on-demand tests. Standard, multiple-choice item formats continue to dominate large-scale computer-based high-stakes testing, resulting in an overreliance on simple, well-structured problems that tap fact retrieval and the use of algorithmic solution procedures.

Classroom instructional uses of assessments.

A distinction has been made between assessments of the outcomes of learning, typically used for grading and accountability purposes (summative assessment), and assessments for learning, used to diagnose and modify the conditions of learning and instruction (formative assessment). The formative use of assessment has been repeatedly shown to significantly benefit student achievement (11, 12). Such effects depend on several classroom practice factors, including alignment of assessments with state standards, quality of the feedback provided to students, involvement of students in self-reflection and action, and teachers actually making adjustments to their instruction based on the assessment results (13).

Technologies are well suited to supporting many of the data collection, complex analysis, and individualized feedback and scaffolding features needed for the formative use of assessment (14). Two illustrative projects, one drawn from science and the other from mathematics, rely on detailed analyses of subject domains and student thinking to provide in-depth assessment and feedback during instruction.

DIAGNOSER is an Internet-based tool that delivers continuous formative assessment and feedback to students and teachers (15, 16). The online assessment identifies problematic "facets" of students' thinking, then provides counterexamples and additional lessons. An example of an item in the DIAGNOSER system is provided in Fig. 2.

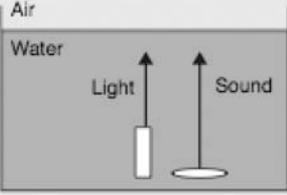
Based on a student's response to this item, as well as others in a set related to this topic area, a diagnosis is done of the student's level of understanding, with feedback provided to both the student and the teacher. In a validation study in a Washington state district, students using DIAGNOSER outperformed their peers on items from the state science test.

ASSISTments is a comprehensive intelligent tutoring system to provide student instruction and support teachers in collection and analysis of their students' data for topics in mathematics (17). Like DIAGNOSER, the system has shown promising initial results and is now being extended to areas of science.

Q7

Chris and Pat think they can explain and predict how waves behave. Pat challenges Chris with the following question.

A flashlight and a sound generator are placed at the bottom of a swimming pool as shown in the diagram. What happens to the speed of the two waves as they move from the water into the air?



Both the sound wave and the light wave speed up in the air.

Both the sound wave and the light wave slow down in the air.

The light wave speeds up; the sound wave slows down.

The speeds do not change because the waves are hitting the surface at 90 degrees.

Continue

Diagnoser: Reflection and Refraction 1

Fig. 2. The DIAGNOSER assessment (15).

Education & Technology

Technology can also support the design of complex, interactive tasks that extend the range of knowledge, skills, and cognitive processes that can be assessed (18). For example, simulations can superimpose multiple representations and permit manipulation of structures and patterns that otherwise might not be visible or even conceivable. Simulation-based assessments can probe basic foundational knowledge such as the functions of organisms in an ecosystem, but, more important, can probe students' knowledge of how components of a system interact as well as abilities to investigate the impacts of multiple variables changing at the same time (19). Moreover, because simulations use multiple modalities and representations, stu-

dents with diverse learning styles and language backgrounds may have better opportunities to demonstrate their knowledge than are possible in text-laden print tests.

In an ongoing program of research and development, WestEd's SimScientists projects are studying the suitability of simulation-based science assessments as summative assessments with the technical quality required for components of an accountability system (19). New SimScientists projects are also studying the use of simulations for curriculum-embedded formative uses of assessment.

Figures 3 and 4 present screen shots of tasks in a SimScientists assessment designed to provide evidence of middle school students' understand-

ing of ecosystems and inquiry practices. Students are presented with the overall problem of preparing a report to describe the ecology of a lake for an interpretive center. They investigate the roles and relationships of the fish and algae in the lake, answering conventional items and also, as shown in Fig. 3, constructing responses such as drawing a food web.

To assess inquiry skills, Fig. 4 shows one of several tasks in which students conduct investigations by manipulating the numbers of fish or plants in a model of the lake and predicting and explaining outcomes. A graph and table provide multiple representations of the population levels. A graph inspector arrow allows students to reexamine the numbers of organisms at different points in time. A camera permits saving each run to a folder for later comparison and analysis.

In a culminating task, students write a report of their findings about the lake. No feedback is presented in the end-of-unit assessment. In a set of embedded assessments, the system identifies types of errors and follows up with increasing levels of feedback and coaching. In the assessment screen shown, after observing animations of the organisms interacting, students draw a food web to depict the flow of matter. An incorrect arrow the student has drawn is highlighted. Levels of feedback and coaching progress from identifying that an error has occurred and asking the student to try again, to explaining the concept (flow of matter), to demonstrating and explaining correct drawing of the arrows.

Research in the SimScientists projects is studying the technical quality of the assessments, the potential of the end-of-unit assessments as components of an accountability system, and the impact of the curriculum-embedded assessments and feedback on student learning. Project designs such as these can document the validity and utility of technology-based assessments for instructional and accountability purposes.

Multilevel State Assessment Systems

It is widely recognized that states must aim for balanced state assessment systems in which classroom, district, and state tests are aligned and mutually reinforcing. The National Research Council (NRC) report *Knowing What Students Know* argued that a balanced assessment system relies on a nested system of assessments that exhibit multiple features (20). One feature is the use of multiple measures covering the full range of standards. Another involves alignment of standards, assessments, curriculum, and instruction within and across different levels of the system (school, district, and state). A third important feature is going beyond annual, on-demand tests to multiple assessments distributed over time combined with timely reporting that offers teachers the opportunity to tailor instruction.

Because of concerns about the adequacy and coherence of state-level assessment systems, the National Science Foundation funded the NRC to offer recommendations to states on the design and

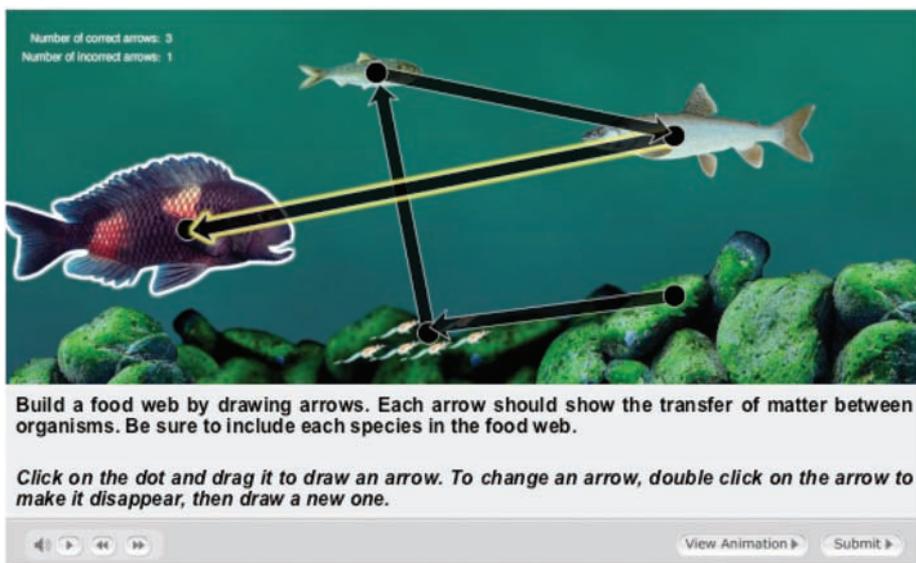


Fig. 3. SimScientists Food Web Construction (19).

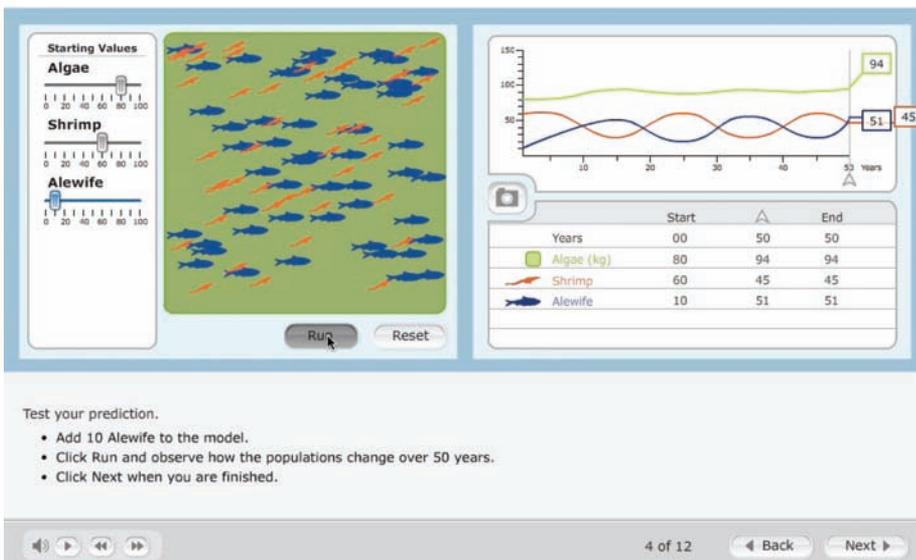


Fig. 4. SimScientists Lake Population Simulation (19).



implementation of their science assessment systems (21). In a report commissioned by that project, Quellmalz and Moody (22) proposed strategies for states to form collaboratives and use technology to create multilevel science assessment systems. With the goal of helping schools and students meet the NCLB goals, states are seeing classroom-based, instructional uses of assessment as a powerful tool for driving student achievement. Such assessment is distinguished from interim assessments administered periodically on a larger scale that are intended to describe the status of student performance after instruction (23).

A key feature in creating a balanced multilevel system is the use of common design specifications that can operate across classroom, district, state, and national levels (22). To enable implementation, online authoring systems are being developed that can assist in creating such common specifications, in streamlining test design, and in reducing development costs (24). Online design systems can also support adaptations of assessments to offer accommodations for special populations while preserving the linkages between targeted standards and designs of the tasks for eliciting evidence of achievement.

Conclusion

Technology helps us do many conventional things in the world of testing and assessment better and faster, and it holds the key to transforming current assessment practice for multiple purposes and at multiple levels ranging from the classroom to state, national, and international levels. We are not there yet, and although many obstacles remain to their widespread use, the next generation of technology-enabled assessments is under development with

several promising cases of design, implementation, and use. Such demonstrations provide a vision of the possible and can help move education toward the design and adoption of more integrated and effective learning-centered assessment tools and systems.

References

1. National Center on Education and the Economy, "Tough choices or tough times: Report of the new commission on the skills of the American workforce" (Jossey Bass, Washington, DC, 2006).
2. National Research Council, *Rising Above the Gathering Storm: Energizing and Employing America for a Brighter Economic Future* (National Academies Press, Washington, DC, 2006).
3. R. E. Bennett, *Educ. Policy Anal. Arch.* **9**, 5 (2001).
4. R. E. Bennett, "Technology for large-scale assessment" (ETS Report No. RM-08-10, Educational Testing Service, Princeton, NJ, 2008).
5. M. Koomen, paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA, April 2006.
6. B. Sandene *et al.*, "Online assessment in mathematics and writing: Reports from the NAEP technology-based assessment project" (NCES 2005-457, U.S. Department of Education, National Center for Educational Statistics, U.S. Government Printing Office, Washington, DC, 2005).
7. T. K. Landauer, D. Laham, P. Foltz, *Assess. Educ.* **10**, 295 (2003).
8. S. Klein, in *Probability and Statistics: Essays in Honor of David A. Freedman*, D. Nolan, T. Speed, Eds. (Institute of Mathematical Statistics, Beachwood, OH, 2008), vol. 2, pp. 76–89.
9. T. Vendlinski, R. Stevens, *J. Technol. Learn. Assessment* **1**, 3 (2002).
10. R. E. Bennett, H. Persky, A. Weiss, F. Jenkins, "Problem solving in technology rich environments: A report from the NAEP technology-based assessment project" (NCES 2007-466, U.S. Department of Education, National Center for Educational Statistics, U.S. Government Printing Office, Washington, DC, 2007).

11. P. Black, D. Wiliam, *Inside the Black Box: Raising Standards Through Classroom Assessment* (King's College, London, 1998).
12. D. Wiliam, in *Second Handbook of Mathematics Teaching and Learning*, F. K. Lester Jr., Ed. (Information Age Publishing, Greenwich, CT, 2007), pp. 1051–1098.
13. P. Black, C. Harrison, C. Lee, B. Marshall, D. Wiliam, *Phi Delta Kappan* **86**, 8 (2004).
14. J. Brown, S. Hinze, J. W. Pellegrino, in *21st Century Education*, T. Good, Ed. (Sage, Thousand Oaks, CA, 2008), vol. 2, chap. 77, pp. 245–255.
15. J. Minstrell, P. Kraus, in *How Students Learn: History, Mathematics, and Science in the Classroom*, J. Bransford, S. Donovan, Eds. (National Academies Press, Washington DC, 2005), chap. 11.
16. A. Thissen-Roe, E. B. Hunt, J. Minstrell, *Behav. Res. Meth. Instrum.* **36**, 234 (2004).
17. M. Feng, N. Heffernan, K. Koedinger, in *Proceedings of the Eighth International Conference on Intelligent Tutoring Systems*, M. Ikeda, K. D. Ashley, T. W. Chan, Eds. (Springer-Verlag, Berlin, 2006), pp. 31–40.
18. E. S. Quellmalz, G. Haertel, "Technology supports for state science assessment systems" (National Research Council, Washington, DC, 2004).
19. E. S. Quellmalz *et al.*, in *Assessing Science Learning: Perspectives from Research and Practice*, J. Coffey, R. Douglas, C. Stearns, Eds. (National Science Teachers Association Press, Washington, DC, 2008), chap. 10.
20. J. Pellegrino, N. Chudowsky, R. Glaser, Eds., *Knowing What Students Know: The Science and Design of Educational Assessment*. (National Academies Press, Washington, DC, 2001)
21. M. R. Wilson, M. W. Bertenthal, Eds., *Systems for State Science Assessment* (National Academies Press, Washington, DC, 2005).
22. E. S. Quellmalz, M. Moody, "Models for multi-level state science assessment systems" (National Research Council, Washington, DC, 2004).
23. M. Perie, S. Marion, B. Gong, "A framework for considering interim assessments" (National Center for the Improvement of Education Assessment, Dover, NH, 2007).
24. R. Mislvey, G. D. Haertel, *Educ. Meas.* **25**, 6 (2006).

10.1126/science.1168046

PERSPECTIVE

Video Games: A Route to Large-Scale STEM Education?

Merrilea J. Mayo

Video games have enormous mass appeal, reaching audiences in the hundreds of thousands to millions. They also embed many pedagogical practices known to be effective in other environments. This article reviews the sparse but encouraging data on learning outcomes for video games in science, technology, engineering, and math (STEM) disciplines, then reviews the infrastructural obstacles to wider adoption of this new medium.

In the 2000-to-2005 time frame, ~450,000 students graduated annually in the United States with a bachelor's degree in STEM (1). These numbers pale in comparison to the reach of a single computer video game (Figs. 1 and 2). *World of*

Warcraft (2), a fantasy game, has over 10 million current subscribers, with ~2.5 million in North America (3). *Food Force* (4), the U.N.-produced game on the mechanics of food aid distribution, saw 1 million players in its first 6 weeks and 4 million players in its first year (5). Additionally, in the realm of K-to-12 science and math education, the virtual world *Whyville* (6), with its game-based

activities, now sports 4 million subscribers (90% North American), with the dominant demographic being 8- to 14-year-old girls (7, 8). Although traditional education institutions pride themselves on educating citizens, they do so at a relatively small scale compared with the media now available. Is it possible to greatly expand the reach of STEM education with the use of video games as the medium? And to what level of effectiveness?

At first, the idea of using video games to teach science and engineering seems laughable. However, sophisticated video game content already exists in topics ranging from immunology (9) (Fig. 3) to numerical methods (10, 11). The examples in Table 1 suggest that video games can yield a 7 to 40% positive learning increase over a lecture program. What's more, there may be additional benefits to poor learners: One variant of the *River City* ecology game (12) diminished the learning gap between D and B students to the point where nearly all students were performing at the B-student level (13).

Learning outcomes are by no means uniformly positive. Results from review studies (14, 15) make

Ewing Marion Kauffman Foundation, 4801 Rockhill Road, Kansas City, MO 64110 USA. E-mail: mmayo@kauffman.org