SOW 06-01-12-(NRC short thought paper).

Embedded Assessment of Informal and Afterschool Science Learning

Diego Zapata-Rivera, Educational Testing Service, Princeton, NJ

Assessing science learning in informal environments involves a series of challenges that are difficult to address using traditional assessment practices (National Research Council, 2009). Some of the assessment challenges inherent in informal and afterschool environments include: (a) interactions in these environments are diverse in terms of duration, type of activity, number of people involved; (b) they usually include emerging behavior due to unpredictable interactions with other participants (e.g., peers, family members, and facilitators); and (c) these environments are characterized by a high degree of freedom and flexibility, which makes it difficult to isolate and measure individual learning. Although traditional measurement instruments have been used to measure learning and engagement in these environments (e.g., self-reports, questionnaires, think-aloud techniques, interviews), these instruments tend to be at odds with the engaging, continuous, and exploratory nature of these environments. This paper describes the potential of employing embedded assessments in the creation of interactive informal learning environments as means of assessing participant learning without disrupting the “flow” experienced by participants (Csikszentmihalyi, 1990).

Embedded Assessment

Embedded assessment is any assessment that is given to students as an integral part of their learning experience. Embedded assessments can be integrated into interactive environments (e.g., simulations, games, and intelligent tutors) at different levels. These levels of integration range from direct assessment activities that may or may not be part of a coherent scenario to completely transparent, unobtrusive sets of actions or “stealth” assessments (Shute, 2011; Shute, Ventura, Bauer, & Zapata-Rivera, 2009).

Embedded assessments gather pieces of evidence that are used to support assessment claims about participants' knowledge and skills. These assessments can be created using Evidence-Centered Design (ECD; Mislevy, Steinberg, & Almond, 2003). ECD is a flexible assessment design methodology that supports the creation of valid assessments by developing evidence-based, argument structures that connect task performance to claims about student knowledge, skills and other attributes (KSAs). ECD makes use of a series of models to make explicit the assessment's evidential argument. These models include: (a) Student or Proficiency Model—describes students' KSAs about which we want to make claims; (b) Evidence Model—describes the relationship between observable outcomes from tasks and the relevant proficiency variables; (c) Task Model—describes the kinds of situations in which we can observe evidence of proficiencies; and (d) Assembly Model—describes the collection of proficiency, evidence, and task models that will constitute a given assessment. It contains the rules used to assemble the form of the assessment seen by a learner from a pool of potential tasks. ECD has been used to design a variety of assessments including assessments embedded in simulations, games, and intelligent tutors (Clarke-Midura, Code, Dede, Mayrath, & Zap, 2011; Shute, et al., 2009; Rupp, Gushta, Mislevy, & Shaffer, 2010; Zapata-Rivera et al., 2007; 2009).

Unobtrusive embedded assessments in informal and afterschool science learning

Stealth assessments are unobtrusive embedded assessments that are woven directly and invisibly into the fabric of the learning or gaming environment. During video game play, students naturally produce rich sequences of actions while performing complex tasks, drawing on the very skills or competencies that we want to assess. The use of stealth assessments does not imply that participants are unaware of data being collected for formative assessment purposes. In fact, learner performance data are usually visible as part of the game (e.g., as performance indicators) and available to different audiences as reports (e.g., reports for teachers or parents). Stealth assessments have been used to assess content (e.g., conceptual physics –static equilibrium; Shute & Kim, 2011) and higher-order thinking skills (e.g., systems thinking, creative problem solving, and causal reasoning; Shute, Masduki, & Donmez, 2010; Shute et al., 2009; Shute & Kim, 2011). Some characteristics of stealth assessments that make them particularly well-suited to address the assessment challenges in informal and afterschool learning environments include: gathering evidence of participants' KSAs in unobtrusive ways, supporting sporadic interactions, and maintaining high levels of participant engagement. Current work on stealth assessment tries to capture unpredictable actions or emerging behavior by applying data-mining techniques to discover interesting patterns of data from participant log files. These patterns are used to refine predefined ECD models. By using a hybrid approach it is possible to draw inferences about student performance based on both student interactions in the game (process data) as well as student responses to embedded assessments. This hybrid approach assembles performance data that naturally emerge through game play (these bottom-up data are mined to identify performance patterns) and performance data from embedded assessments that focus on the KSAs of the ECD model (top-down data).

Simulations and video games comprise an important category of informal science learning environments. Video games have been used as part of an exhibit or as a mechanism for extending the museum experience before and after the visit to the museum (e.g., by playing relevant video games as part of an after-school program). Simulations and video games can be used to support meaningful activities, learning that takes place invisibly and naturally, and social interaction and discussion (Norman, 2001). The majority of teenagers in the United States play video games (Lenhart et al., 2008). Researchers have explored the use of video games for supporting student learning of valued KSAs in informal and formal environments (Barab et al., 2010; Dede, 2007; National Research Council, 2011; Shaffer, 2006; Squire & Jenkins, 2003). Stealth assessments have the potential for providing valid information about what students are learning when playing video games.

Assessing participants’ learning across diverse informal and formal learning environments is an ongoing challenge. Advances in mobile and adaptive technologies make it possible to think about lifelong learner models that are controlled by the learner and could be shared across platforms and systems (Kay, 2008). This learner model can “remember” user interactions in a variety of contexts. Applications that share the learner model can use this information to adapt their interaction to the knowledge, skills, needs, or preferences of the learner. Pieces of evidence (e.g., users’ interactions) collected by these applications in various contexts (in and out of school) can be integrated into a distributed learner model[1] (Assad, Carmichael, Kay & Kummerfeld, 2007; Brusilovsky, 2004; Zapata-Rivera & Greer, 2004), which can be used to create personalized experiences. Museum exhibits can use this lifelong learner model to create a personalized experience that builds on the contents of the model (e.g., “remembering” the interactions of a returning visitor to create a new experience; Kuflik, Kay, & Kummerfeld, 2010). DiCerbo and Behrens (2012) describe the concept of “an assessment ecosystem” as an environment in which information is accumulated from a variety of natural digital experiences to form a cohesive view of students’ knowledge, skills, and attributes. Technologies that can facilitate this vision include ECD, “big data” technologies (White, 2009), and Bayesian networks (Almond, DiBello, Moulder, & Zapata-Rivera, 2007; Levy & Mislevy, 2004; Pearl, 1988).

Quality of embedded assessments

[1] A distributed learner model refers to a model of the learner that is created, maintained and used by a variety of applications in a distributed learning environment.

SOW 06-01-12-(NRC short thought paper).

Research has begun to evaluate the use of simulations for the purpose of science assessment. This research provides some evidence of valid and reliable use of simulation-based assessments in formal learning environments (e.g., Bennett, Persky, Weiss, & Jenkins, 2007; Quellmalz et al., 2011). These assessments have been integrated successfully into traditional computer-based summative and formative assessments. The interactive nature of and use of multiple media in simulations and games makes accessibility and fairness issues a difficult challenge. These issues are addressed by a combination of applying extensions of ECD (Hansen & Mislevy, 2008) and universal design principles (Thompson et al., 2002).

While research on the use of simulations for assessment in formal learning environments yields promising evidence, the same amount of evidence is not yet available for the use of stealth assessments in games. Several studies exploring validity and reliability issues of stealth assessments are underway (Shute & Ventura, 2011). Work on accessibility and fairness issues of stealth is not available yet. Zapata-Rivera and Bauer (2011) discuss some of the challenges relating to the implementation of valid, reliable, and fair assessment in games. These challenges include the following:

- *Introduction of construct irrelevant content and skills.*[2] The authenticity added by the context of a game may also elicit irrelevant knowledge, skill, or other attributes (Messick, 1994). When designing interactive gaming activities it is easy to introduce content and interactions that impose requirements on KSAs that are unrelated to the construct of interest. For each game element, designers must consider whether those elements lead students to use knowledge and skills related to the goals of the assessment.

[2] According to Messick, construct irrelevant variance appears when the “test contains excess reliable variance that is irrelevant to the interpreted construct” (Messick, 1989).

- *Demands on working memory*. Related to construct irrelevant variance is the issue of the demands that game-like assessments place upon students' working memory. By designing assessments with higher levels of interactivity and engagement, it is easy to increase cognitive processing demands to the point of reducing measurement quality. There are many design strategies that can be used to address cognitive load in the design of game-like assessments. For example, Mayer and Moreno (2003) propose a dual processing theory for images and sounds in multimedia learning systems and provide nine research-based principles to reduce cognitive load.

- *Accessibility issues*. Games that make use of rich, immersive graphical environments (e.g., sophisticated navigation controls) impose visual, motor, auditory, and other demands on the player, which create accessibility issues for students with disabilities. However, creating environments that do not make use of some of these technological advances (e.g., a 3D immersive environment) may reduce student engagement, especially for students who are used to interacting with highly interactive games. One potential solution to this issue is to construct parallel environments that do not impose the same visual, motor, and auditory demands but still assess the constructs of interest.

- *Tutorials and familiarization*. Although the majority of students have played some sort of video game in their lives, students will need support to understand how to navigate and interact with the graphical environment for any particular assessment. Lack of familiarity with navigation controls may negatively influence student performance and student motivation (e.g., Lim, Nonis, & Hedberg, 2006). The use of tutorials and demos can support this familiarization process. The tutorial can also be used as an engagement element (e.g., Armstrong & Georgas, 2006).

- *Type and amount of feedback*. Embedded assessments can be used to provide feedback to learners. Feedback is a key component of formative assessments. Research shows that interactive computer applications that provide immediate, task-level feedback to students can positively contribute to student learning (e.g., Hattie & Timperley, 2007; Shute, 2008). Depending on the purpose of the assessment (i.e., formative or summative) different types of feedback need to be available. Immediate feedback that results from a direct manipulation of objects in the game can provide useful information to guide exploration or refine interaction strategies. Availability of feedback may influence motivation and the quality of the evidence produced by the system. Measurement models need to take into account the type of feedback that has been provided to students when interpreting the data gathered during their interaction with the assessment system.
- *Interaction issues (re-playing, number of attempts and revisions).* Allowing for various attempts or revisions while providing immediate feedback has implications for evidence gathering and evidence accumulation processes. As in the case of feedback, measurement models need to handle the number of attempts and revisions. This could be done by comparing the outcomes of consecutive actions/events or by interpreting a subset of actions including the type of feedback received. Based on the type of assessment, operational constraints (e.g., time) may impose a limit on the number of attempts allowed before moving to the next scenario.
- *Handling dependencies among actions*. Dependencies among actions/events can be complex to model and interpret. Assumptions of conditional independence required by some measurement models may not hold in complex interactive scenarios. Designing scenarios carefully in order to minimize dependencies will help reduce the complexity of

measurement models. Using data mining techniques to support evidence identification can also help with this issue.

We expect that research focused on these issues will shed light on the strengths and limitations of embedding stealth assessments within games that could be played in formal or informal learning environments. This research may also be able to inform embedded assessment designs outside of computer game environments, by unearthing general principles that need to be considered in order to ensure that embedded assessments measure what we care about. Many of the bullet points listed above, for instance, would equally apply to design of a science museum manipulative or hands-on interactive activity. An interactive activity could be designed to unobtrusively measure a museum visitor's level of understanding of the underlying concepts and to respond with coaching or feedback that would enhance the visitor's experience.

Conclusions

Embedded assessments and related technologies have the potential to contribute to assessing science learning in informal and afterschool environments. Current results seem encouraging. However, more research is needed. Exploring the strengths and limitations of applying these new types of assessments in informal and afterschool science learning requires a multidisciplinary team of researchers and practitioners involving experts in areas such as video games, education, cognitive science, informal education, and measurement.

This research may also be able to inform embedded assessment designs outside of computer game environments, by unearthing general principles that would help to ensure that embedded assessments measure what we care about.

SOW 06-01-12-(NRC short thought paper).

References

Almond, R. G., DiBello, L. V., Moulder, B., & Zapata-Rivera, J. D. (2007). Modeling diagnostic assessments with Bayesian networks. *Journal of Educational Measurement*, 44, 341-359.

Armstrong, A., & Georgas, H. (2006). Using interactive technology to teach information literacy concepts to undergraduate students, *Reference Services Review*, 34 (4), 491 – 497.

Assad, M., Carmichael, D.J., Kay, J., and Kummerfeld, B. (2007). PersonisAD: Distributed, Active, Scrutable Model Framework for Context-Aware Services. *Proc. Fifth Int'l Conf. Pervasive Computing (PERVASIVE '07)*, pp. 55-72.

Barab, S. A., Dodge, T., Ingram-Goble, A., Pettyjohn, P., Peppler, K, Volk, C., & Solomou, M. (2010). Pedagogical dramas and transformational play: Narratively rich games for learning. *Mind, Culture, and Activity*, 17(3), 235-264.

Bennett, R. E., Persky, H., Weiss, A., & Jenkins, F. (2007). *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.

Brusilovsky, P. (2004). KnowledgeTree: A Distributed Architecture for Adaptive e-Learning. *Proc. 13th Int'l World Wide Web Conf. Alternate Track Papers & Posters*, pp. 104-113.

Clarke-Midura, J., Code, J., Dede, C., Mayrath, M., & Zap, N. (2011). Thinking outside the bubble: Virtual performance assessments for measuring complex learning. In M.C. Mayrath, J. Clarke-Midura, & D. Robinson (Eds.), *Technology-based assessments for 21st century skills:*

SOW 06-01-12-(NRC short thought paper).

Theoretical and practical implications from modern research. Charlotte, NC: Information Age. 125-147

Csikszentmihalyi, M. (1990) Flow: The psychology of optimal experience. Harper and Row.

DiCerbo, K. & Behrens, J. (2012). From Technology-Enhanced Assessment to Assessment-Enhanced Technology. Paper presented at Paper presented at the annual meeting of the National Council on Measurement in Education (NCME), Vancouver, BC. Canada.

Dede, C. (2007). Transforming education for the 21st Century: New pedagogies that help all students attain sophisticated learning outcomes. Harvard University: Commissioned by the NCSU Friday Institute. Retrieved on April 20, 2012 from http://tdhahwiki.wikispaces.com/file/view/Dede_21stC-skills_semi-final.pdf

Hattie, J., and Timperley, H. (2007). The power of feedback. *Review of Educational Research*. 77, 1, 81-112.

Hansen, E. G., & Mislevy, R. J. (2008). Design patterns for improving accessibility for test takers with disabilities (ETS Research Rep. No. RR-08-49). Princeton, New Jersey: ETS.

Kay, J. (2008) Lifelong Learner Modeling for Lifelong Personalized Pervasive Learning. IEEE Transactions on Learning Technologies, Vol. 1, No. 4, pp. 215-227.

Kuflik, T., Kay, J., and Kummerfeld, B. (2010). Lifelong Personalized Museum Experiences. *In proceedings of Pervasive User Modeling and Personalization PUMP10* at UMAP2010.

Lenhart, A., Kahne, J., Middaugh, E., Macgill, A., Evans, C., & Vitak, J. (2008). Teens, Video Games and Civics. Pew Internet & American Life Project. Available at: http://pewinternet.org/PPF/r/263/report_display.asp.

SOW 06-01-12-(NRC short thought paper).

Levy, R., & Mislevy, R. J (2004). Specifying and refining a measurement model for a computer based interactive assessment. *International Journal of Testing*, 4, 333-369.

Lim, C. P., Nonis, D., & Hedberg, J. (2006). Gaming in a 3-D multiuser virtual environment: engaging students in Science lessons. *British Journal of Educational Technology*, 37(2), 211-231

Mayer, R.E., and Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist*, 38, (1), 43-52.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement*, (3rd ed.). Washington, D.C.: American Council on Education.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, 23(2), 13-23.

Mislevy, R.J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessment (with discussion). Measurement: Interdisciplinary Research and Perspective, 1, 3–62.

National Research Council (2009). *Learning Science in Informal Environments: People, Places, and Pursuits. Committee on Learning Science in Informal Environments*. Philip Bell, Bruce Lewenstein, Andrew W. Shouse, and Michale A. Feder, Editors. Board on Science Education, Center for Education. Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

National Research Council (2011). *Learning Science Through Computer Games and Simulations*. Committee on Science Learning: Computer Games, Simulations, and Education, Margaret A. Honey and Margaret Hilton, Editors. Board on Science Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

SOW 06-01-12-(NRC short thought paper).

Norman, D. (2001). The Future of Education: Lessons Learned from Video Games and Museum Exhibits. Available at: http://jnd.org/dn.mss/the_future_of_education_lessons_learned_from_video_games_and_museum_exhibits.html

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann Publishers.

Quellmalz, E. S., Timms, M. J., Buckley, B. C., Davenport, J., Loveland, M., & Silberglitt, M. D. (2011). 21st Century Dynamic Assessment. In M.C. Mayrath, J. Clarke-Midura, & D. Robinson (Eds.), *Technology-based assessments for 21st century skills: Theoretical and practical implications from modern research*. Charlotte, NC: Information Age. 55-90.

Rupp, A.A., Gushta, M., Mislevy, R.J., & Shaffer, D.W. (2010). Evidence-centered design of epistemic games: Measurement principles for complex learning environments. *Journal of Technology, Learning, and Assessment*, 8(4). Retrieved from http://www.jtla.org.

Shaffer, D. W. (2006). *How computers help children learn*. Basingstoke: Palgrave Macmillan.

Shute, V. (2008) Focus on Formative Feedback. *Review of Educational Research*. 78 (1), 153-189.

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.

Shute, V. J., Masduki, I., & Donmez, O. (2010). Conceptual framework for modeling, assessing, and supporting competencies within game environments. *Technology, Instruction, Cognition, and Learning*, 8(2), 137-161.

SOW 06-01-12-(NRC short thought paper).

Shute, V. J., & Kim, Y. J. (2011). Does playing the World of Goo facilitate learning? In D. Y. Dai (Ed.), *Design research on learning and thinking in educational settings: Enhancing intellectual growth and functioning* (pp. 359-387). New York, NY: Routledge Books.

Shute, V. J., Ventura, M. (2011). Digital Games, Learning, and Stealth Assessment. Available at: http://myweb.fsu.edu/vshute/CPDoverview.pdf

Shute, V. J., Ventura, M., Bauer, M. I., & Zapata-Rivera, D. (2009). Melding the power of serious games and embedded assessment to monitor and foster learning: Flow and grow. In U. Ritterfeld, M. J. Cody, & P. Vorderer (Eds.), *Serious Games: Mechanisms and Effects*. Philadelphia, PA: Routledge/LEA. 295-321.

White, T. (2009). *Hadoop: The Definitive Guide*. O'Reilly Media

Squire, K., & Jenkins, H. (2003). Harnessing the power of games in education. Retrieved on November 20, 2009 from http://website.education.wisc.edu/kdsquire/manuscripts/insight.pdf

Thompson, S. J., Thurlow, M. L., Quenemoen, R. F., & Lehr, C. A., (2002). Access to computer-based testing for students with disabilities (Synthesis Report 45). National Center on Educational Outcomes, University of Minnesota, Minneapolis, MN . Retrieved from education.umn.edu/NCEO/OnlinePubs/Synthesis45.html

Zapata-Rivera, D. & Bauer, M. (2011) Exploring the Role of Games in Educational Assessment. In M.C. Mayrath, J. Clarke-Midura, & D. Robinson (Eds.), *Technology-based assessments for 21st century skills: Theoretical and practical implications from modern research*. Charlotte, NC: Information Age. 149-172

Zapata-Rivera, D. and Greer, J. (2004). Inspectable Bayesian Student Modelling Servers in Multi-Agent Tutoring Systems. *International Journal of Human-Computer Studies*, vol. 61, no. 4, pp. 535-563, 2004.

SOW 06-01-12-(NRC short thought paper).

Zapata-Rivera, D., VanWinkle, W., Doyle, B., Buteux, A., & Bauer, M. I. (2009). Combining Learning and Assessment in Assessment-based Gaming Environments: A Case Study from a New York City School. *Interactive Technology and Smart Education*. Vol, 6, 3. 173 – 188. Emerald Group Publishing Limited.

Zapata-Rivera, D., Vanwinkle, W., Shute, V., Underwood, J., & Bauer, M (2007). English ABLE. In *Artificial Intelligence in Education - Building Technology Rich Learning Contexts That Work*. Luckin, R., Koedinger, K., & Greer, J. (Eds.) Vol. 158. 323 – 330.