

What Do We Know About Assessment in Games?

**Eva L. Baker and Girlie C. Delacruz
University of California, Los Angeles**

**National Center for Research on Evaluation, Standards, and Student Testing
(CRESST)**

This presentation will address issues surrounding assessment in games. We are focused on the measurement of outcomes or the attainment of specific levels of proficiency in cognitive and procedural tasks. We are also interested in understanding and instrumenting the process of learning in a manner somewhat more sophisticated and much more usable than counting clicks. Along with embedded measures of attention and engagement, these measures included in a game can permit the inference to be made about the effects of the game on learning. We use the term assessment here to be more congruent with the idea of measurement than the notion of evaluation. We see measurement as providing the key measures that are the backbone of an evaluation study, but may be insufficient in the evaluation of the game or system in which the game is embedded. Other features of an evaluation would include the selection criteria for players, their persistence, the cost of the game compared to other methods of achieving goals, and niceties involving retention, time to refresh skill or knowledge decay, and feasibility and perception matters.

Educators and trainers have recognized the potential of computer games for education and training since the 1970s (Donchin, 1989; Malone, 1981; Malone & Lepper, 1987; Ramsberger, Hopwood, Hargan, & Underhill, 1983; Ruben, 1999; Thomas & Macredie, 1994). A renewed interest and optimism have emerged around games, particularly games that incorporate highly interactive multimedia (e.g., Dickey, 2005; Gee, 2003, 2004; Kafai, 2006; Kafai, Franke, Ching, & Shih, 1998; Klopfer & Squire, 2004; O'Neil & Perez, 2008; Shaffer, Squire, Halverson, & Gee, 2005). Others caution that a lack of instructional design and the inclusion of “bells and whistles”—intended to be motivating to students—will distract students from attaining desired outcomes (e.g., Clark, 2003; Kalyuga, Chandler, Touvinen & Sweller, 2001; Mayer, 2004; van Merriënboer, Clark, & de Croock, 2002).

The proliferation of games and platforms presents an interesting opportunity: To what extent can learning opportunities be extended beyond formal educational settings? Increasing time-on-task may be particularly important for students at risk for failing. By definition such opportunities will be unsupervised and students need to have strong self-regulation skills (O'Neil, 2002). To be effective in unsupervised settings, learners need to

be able to assess their own performance, detect knowledge and skill gaps, and seek the appropriate help when needed.

The allure of using games for learning is their potential to provide multiple benefits: complex and diverse approaches to learning processes and outcomes; high interactivity and engagement; immediate feedback; enjoyment; intense engagement (flow); problem solving; scaffolding; user control; adaptive challenge; contextual learning; ability to address cognitive as well as affective learning issues; and perhaps most importantly, motivation for learning (de Freitas, 2006; Kirriemuir & McFarlane, 2003; O'Neil, Wainess, & Baker, 2005; Wideman, Owston, Brown, Kushniruk, Ho, & Pitts, 2007). Games have the potential in immersive environments to help students develop cognitive readiness, and self-monitoring skills and to engage with content in greater depth and sophistication than the typical classroom experience allows. Gaming environments can provide students opportunities to explore complex worlds, ponder physical, social, and ethical dilemmas, and witness the implications of their actions as they play out in real time and often in unpredictable ways.

The real question is “how do you know?” “...Such stuff as dreams are made...” Shakespeare, William, *The Tempest*, Act 4, scene 1, lines 148–158). Whether games can be effective is unclear as there is not a large empirical base to draw on to make a firm conclusion (Federation of American Scientists, 2006; Fletcher & Tobias, 2006; de Freitas, 2006; Kirriemuir & McFarlane, 2003; Mitchell & Savill-Smith, 2004; O'Neil, Wainess, & Baker, 2005; Randel, Morris, Wetzell, & Whitehill, 1992; Wideman et al., 2007). However, like all other learning environments, especially technology-based ones, the issue of good assessment practices is critical for effective learning in games (Baker & Delacruz, 2008; Baker, Chung, & Delacruz, 2007). If a game is to maximize knowledge and skill acquisition, retention, and transfer in short periods of time, within and outside of supervised classroom settings, assessments embedded in the game become necessary to develop a high-functioning, highly effective implementation. These assessments or measures would consist of precursors, subtasks, brush-ups, and criterion tasks at various levels of challenge and proficiency. The assessments also give immediate feedback both to the student about his or her ongoing performance, and either on-the-fly or summary feedback to the instructor or teacher for instructional purposes or classification purposes.

Typically, the game's scoring mechanism, such as the number of contacts made, obstacles overcome, and so on, is tallied against a time bar. This provides the sense of an evaluation of performance, and when summarized, scores can be inferred to represent something, like a competitive position in a distribution of players or proficiency along

some continuum of performance. However, what is counted rarely relates directly to the grander cognitive or procedural aspects promoted for the game. Rather, these metrics are developed for motivational purposes (e.g., provide just the right amount of challenge), not for evaluative ones.

Other approaches to game-based assessment beyond simple scoring mechanisms have involved the idea of either adding on tasks to a game (e.g., wrap-around assessments) or allowing a toggle opportunity to stop the game world as it progresses in order to ask relevant questions. These approaches typically have the merit of trying to assess a particular knowledge, skill, or attribute that the game addresses. However, the technique is selected not because of measurement merit but out of desperation.

The promise of game-based assessment is the potential for games to be potent formative assessment tools (Baker & Delacruz, 2008; Young, 1995). In use, embedded assessments provide the trigger for changes in options for the learner, and ideally, the assessments would be experienced as part of the game, providing hidden modulation of the challenges related to the academic material and cognitive demands that are necessary for sustained play. The information about how students are doing during the game needs to be provided to the teacher or the students themselves and ideally throughout game play. We can capture this process data through evaluation of students' online clickstream behavior to support inferences about students' ongoing understanding. The use of performance and process information, by the student or the teacher, with the intent to improve learning underlies the idea of formative assessment (Airasian & Jones, 1993; Baker, 1974; Black & Wiliam, 1998; Markle, 1967; National Research Council, 2001, 2003; Scriven, 1967). Reviews of research on the topic (Black & Wiliam, 1998; Crooks, 1988; Kluger & DeNisi, 1996; Natriello, 1987; Nyquist, 2003; Wiliam, Lee, Harrison, & Black, 2004) report relatively high effect sizes when feedback was used in combination with adaptation of instruction.

This may be one of the biggest potentials for game design—the ability of process data to help explain learning outcomes (e.g., use of productive or unproductive strategies), sense and adapt to students' evolving understanding of the domain, misconceptions, or gaps in domain knowledge. However, while it would be ideal to make assessments that do not disturb the flow of game play, complex 3-D games rarely make available source code to use for seamless assessment capabilities, for well-known reasons, including proprietary interests, corruption of code, and concerns about the outcomes of the assessment itself.

Therefore, for games to be effective in an educational context, they have to be integrated into the curriculum or training sequence at the outset of their design, rather than as an add-on (de Freitas, 2006; Mitchell & Savill-Smith, 2004; Sandford, Ulicsak, Facer, & Rudd, 2006). At the outset, we must design the assessment architecture, in other words, embed the assessment in the transactions of the game and build it into a game's underlying engine. Maximizing the potential of process data requires a tight (conceptual) coupling among the set of goals of learning, purpose of assessment, student behaviors, student responses, task design, and assessment design. Paired with analysis of cognitive demands (Baker & Delacruz, 2008) and generated by subject matter experts, the architecture can be developed into an ontology which will capture the necessary information and relationships that will serve as the basis for our game design (Chung, Delacruz, & Bewley, 2004; Chung, Delacruz, Dionne, & Bewley, 2003).

To do so requires first specifying the range of learning outcomes of interest. Second, decisions in advance should be made about the evidence needed to support any inferences made about performance. Whether it is to certify proficiency or determine necessary gaps for remediation, what is captured and how the game responds to the behaviors must be systematically chosen to align with the learning objectives of the game and the purposes it will serve. The difficulty here is to meld this architecture with the instructional learning system as well as to assure that various scenarios to which the player is exposed provide comparable or comprehensive attention to the performance expectations. Unless these elements are tentatively structured at the front-end, there is little hope that the game design (e.g., the game's rules, response expectations of the players, or narrative) will support the types of learning to be experienced or the framework for their evaluation (Baker & Delacruz, 2008).

Moreover, this problem has a strong social component, for the values and expertise of the content experts, narrative specialists, software designers, skilled assessors, and learning specialists must be understood, accepted, and respected. This synergistic and multi-disciplinary approach to game design will effectively maximize each component of the game. In other words, a strong sense of collective efficacy needs to be developed (Baker, Niemi, & Chung, 2008; Hargreaves, 2001, 2004) and a compendium of related assessment purposes should be considered.

Validation Issues

For the most part, performance-focused games still need to confront the question of the validity of the outcomes. Beyond the entertainment value, which can be assessed in

numbers of downloads, purchases, connect time, and reported happy feelings (or even sadness as one of our colleagues has expressed at his inability to disengage after many consecutive hours of play), we must come to the topic of reliability, fairness, and validity. As in many new media, the quality of the measures are assumed to be fine if they map back to the content of the implementation, or if they look like other forms of tests. However, our game-based assessments must also be fair and accurate without inadvertent barriers or support given to different groups of learners (Messick, 1989) and they must be valid for the full range of purposes for which they are intended (American Educational Research Association, American Psychological Association, and National Council for Measurement in Education, *Standards for Educational and Psychological Testing*, 1999). Game performance impact should be judged in the relevant high-stakes setting it may be targeted towards (i.e., job performance, test performance, or project design used for evaluative purposes). The assessment must also be stable so that we can trust the results given by the assessment.

Given that assessment validity is not a property of the test, but rather related to how the information garnered from the assessment is to be used and interpreted, one must gather the appropriate technical evidence to support the inferences drawn about performance and make it readily available for critique and replication (AERA, APA, & NCME, 1999). The publication of technical quality data on game performance needs to be as regular and available as system requirements for the game. It should specify the ideal subset of skills for the would-be player, the slope of typical acquisition of various outcomes over time, the levels of accomplishment that can be attained with benchmarks to real people or situations provided, evidence of the degree to which the game promotes the cognitive, content, affective, and entertainment goals it purports to, and the interaction of the game with its setting, that is, how well it performs with and without supervision, and what kind of expert assistance is needed outside the game, etc. This explicit specification provides the necessary rationale for the interpretation and use of the outcomes of the game-based assessments. This is particularly important when games are used for high-stakes evaluation of performance. When such data are routinely exposed or expected, then game designers will move toward a design approach that takes the consumer (or the purchasing power) seriously, and much of the hand-wrenching difficulties now in place for quality game assessment will magically disappear.

In conclusion, what do we propose are the key elements to consider to successfully incorporate good assessment practices to make effective game-based learning?

1. *Assessment architecture*: Game-based assessment, whether within the target game, an external test, or as a supplemental gaming situation, needs an architecture. What is most important in the architectural aspects of the design are the creation of expectations for the depth and breadth of sampling performance, and metrics for the evaluation of the performance, best developed by inferences from expert performance than only from designations by experts.
2. *Developmental assessment* to provide explicit design information for creators and validators of games (Markle, 1967). This requires making the design elements available with enough specificity that it can serve a practical function for future creators as well as a technical function for those who wish to evaluate the game.
3. *Formative assessment* for users to give information for immediate feedback to the student and teacher. Typically, one would wish to provide guidance to the learner in relatively small chunks near the beginning of a game-playing cycle to keep them on track, within, of course, the bounds of exploration and other “fun” failures planned for the player. Later, however, the formative feedback might well be scheduled in different, longer chunks, or require higher levels of complex performance in order to kick in.
4. *Self-assessment measures* to promote self-regulated learning to support the unsupervised operation of games (e.g., in homework settings), and to create sustained usage and the skills for independent learning.
5. *Criterion assessment* where the outcomes of the game are measured (either immediately or delayed in time) against various targeted levels of proficiency.
6. *Transfer and generalization* where the outcomes of the game are assessed under different conditions (e.g., format differences such as simulations or standard tests) and different constraints (e.g., requirements to find available knowledge, variable time limits, different distractions).

Together these elements are key to ensuring that assessments are utilized appropriately and effectively in games. They require that the goals for assessment and learning be considered concurrently rather than as an afterthought. Finally, we cannot stress enough the importance of making explicit with enough specificity the elements of the game design in order for proper integration, utilization, and validation.

References

- American Educational Research Association (AERA), American Psychological Association (APA), and National Council for Measurement in Education (NCME). (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Airasian, P. W., & Jones, A. M. (1993). The teacher as applied measurer: Realities of classroom measurement and assessment. *Applied Measurement in Education*, 6, 241-254.
- Baker, E. L. (1974). Formative evaluation of instruction. In W. J. Popham (Ed.), *Evaluation in education: Current applications* (pp. 531-585). Berkeley, CA: McCutchan. (ERIC Document Reproduction Service No. ED 123 239).
- Baker, E. L., Chung, G. K. W. K., & Delacruz, G. C. (2007). Design and validation of technology-based performance assessments. In J. M. Spector, M. D. Merrill, J. J. G. van Merriënboer, & M. P. Driscoll (Eds.), *Handbook of research on educational communications and technology* (3rd ed., pp. 595-604). Mahwah, NJ: Erlbaum.
- Baker, E. L., & Delacruz, G. C. (2008). A framework for the assessment of learning games. In H. F. O'Neil & R. S. Perez (Eds.), *Computer games and team and individual learning* (pp. 21-37). Oxford, UK: Elsevier.
- Baker, E. L., Niemi, D., & Chung, G. K. W. K. (2008). Simulations and the transfer of problem-solving knowledge and skills. In E. L. Baker, J. Dickieson, W. Wulfeck, & H. F. O'Neil (Eds.), *Assessment of problem solving using simulations* (pp. 1-17). New York: Erlbaum.
- Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy, and Practice*, 5(1), 7-74.
- Chung, G. K. W. K., Delacruz, G. C., & Bewley, W. L. (2004). Performance assessment models and tools for complex tasks. *International Test and Evaluation Association (ITEA) Journal*, 25(1), 47-52.
- Chung, G. K. W. K., Delacruz, G. C., Dionne, G. B., & Bewley, W. L. (2003). Linking assessment and instruction using ontologies. *Proceedings of the IITSEC*, 25, 1811-1822.
- Clark, R. E. (2003). Fostering the work motivation of teams and individuals. *Performance Improvement*, 42(3), 21-29.
- Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research*, 58(4), 438-481.
- de Freitas, S. (2006). *Learning in immersive worlds: A review of game-based learning*. London: Joint Information Systems Committee (JISC).
- Donchin, E. (1989). The learning strategies project. *Acta Psychologica*, 71, 1-15.
- Dickey, M. D. (2005). Engaging by design: How engagement strategies in popular computer and video games can inform instructional design. *Educational Technology Research & Development*, 53(2), 67-83.
- Federation of American Scientists (FAS). (2006). *Harnessing the power of video games for learning*. Washington, DC: Author.
- Fletcher, J. D., & Tobias, S. (2006) Using games and simulations for instruction: A research review. In, *Proceedings of the New Learning Technologies 2006 Conference*, Warrenton, VA: Society for Applied Learning Technology.

- Gee, J. P. (2003). *What video games have to teach us about learning and literacy*. New York: Palgrave/Macmillan.
- Gee, J. P., (2004). Learning by design: Games as learning machines. *Interactive Educational Multimedia*, 8, 15-23.
- Hargreaves, D. H. (2001). A capital theory of school effectiveness and improvement. *British Educational Research Journal*, 27(4), 487-503.
- Hargreaves, D. H. (2004). *Learning for life: The foundations of lifelong learning*. Bristol, UK: Policy Press.
- Kafai, Y. B. (2006). Playing and making games for learning: Instructionist and constructionist perspectives for game studies. *Games and Culture*, 1, 36-40.
- Kafai, Y. B., Franke, M., Ching, C., & Shih, J. (1998). Game design as an interactive learning environment fostering students' and teachers' mathematical inquiry. *International Journal of Computers for Mathematical Learning*, 3, 149-184.
- Kalyuga, S., Chandler, P., Touvinen, J., & Sweller, J. (2001). When problem-solving is superior to worked examples. *Journal of Educational Psychology*, 93, 579-588.
- Kirriemuir, J., & McFarlane, A. E. (2003) *Literature review in games and learning, Report 8*, Bristol: Nesta Futurelab.
- Klopfer, E., & Squire, K. (2004). Getting your socks wet: Augmented reality environmental science. In *Proceedings of the 6th international conference on the learning sciences (ICLS)* (p. 614), Los Angeles, CA. Retrieved October 29, 2007 from <http://portal.acm.org/citation.cfm?id=1149238>
- Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, 119, 254-284.
- Malone, T. W. (1981). Toward a theory of intrinsically motivating instruction. *Cognitive Science*, 4, 333-369.
- Malone, T. W., & Lepper, M. R. (1987). *Aptitude, learning and instruction III: Cognitive and affective process analysis*. Hillsdale, NJ: Erlbaum.
- Markle, S. M. (1967). Empirical testing of programs. In P. C. Lange (Ed.), *Programmed instruction* (The sixty-sixth yearbook of the National Society for the Study of Education, Part II, pp. 104-138). Chicago: National Society for the Study of Education.
- Mayer, R. E. (2004). Should there be a three-strikes rule against pure discovery learning? *American Psychologist*, 59, 14-19.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan.
- Mitchell, A., & Savill-Smith, C. (2004). *The use of computer and video games for learning. A review of the literature*. London: Learning and Skills Development Agency.
- National Research Council. (2001). *Classroom assessment and the National Science Education Standards. Committee on Classroom Assessment and the National Science Education Standards*. J. M. Atkin, P. Black, & J. Coffey (Eds.). Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.
- National Research Council. (2003). *Assessment in support of instruction and learning: Bridging the gap between large-scale and classroom assessment*. Workshop

- report. Committee on Assessment in Support of Instruction and Learning. Board on Testing and Assessment, Committee on Science Education K-12, Mathematical Sciences Education Board. Center for Education. Division of Behavioral and Social Sciences and Education. Washington, DC: National Academies Press.
- Natriello, G. (1987). The impact of evaluation processes on students. *Educational Psychologist*, 22(2), 155-175.
- Nyquist, J. B. (2003, December). *The benefits of reconstructing feedback as a larger system of formative assessment: A meta-analysis*. Unpublished master's thesis, Vanderbilt University, Nashville, TN.
- O'Neil, H. F., Jr. (2002). Perspectives on computer-based assessment of problem solving [Special issue]. *Computers in Human Behavior*, 18, 605-607.
- O'Neil, H. F., & Perez, R. S. (2008). *Computer games and team and individual learning*. Oxford, UK: Elsevier.
- O'Neil, H. F., Wainess, R., & Baker, E. L. (2005). Classification of learning outcomes: Evidence from the computer games literature. *The Curriculum Journal*, 16(4), 455-474.
- Ramsberger, P. F., Hopwood, D., Hargan, C. S., & Underhill, W. G. (1983). *Evaluation of a spatial data management system for basic skills education. Final Phase I Report for Period 7 October 1980- 30 April 1983* (HumRROFR-PRD-83-23). Alexandria, VA: Human Resources Research Organization.
- Randel, J. M., Morris, B. A., Wetzel, C. D., & Whitehill, B. V. (1992). The effectiveness of games for educational purposes: A review of recent research. *Simulation & Gaming*, 23, 261-276.
- Ruben, B. D. (1999). Simulations, games, and experience-based learning: The quest for a new paradigm for teaching and learning. *Simulation & Gaming*, 30, 498-505.
- Sandford, R., Ulicsak, M., Facer, K., & Rudd, T. (2006). *Teaching with games: Using commercial off-the-shelf computer games in formal education*. Bristol, England: Futurelab. Retrieved October 29, 2007 from http://www.futurelab.org.uk/projects/teaching_with_games/research/final_report/
- Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler, R. M. Gagne, & M. Scriven (Eds.), *Perspectives of curriculum evaluation* (American Educational Research Association Monograph Series on Curriculum Evaluation, No. 1, pp. 39-83). Chicago: Rand McNally.
- Shaffer, D. W., Squire, K. D., Halverson, R., & Gee, J. P. (2005). Video games and the future of learning. *Phi Delta Kappan*, 87, 104-111.
- Thomas, P., & Macredie, R. (1994). Games and the design of human-computer interfaces. *Educational Technology*, 31(2), 134-142.
- van Merriënboer, J. J. G., Clark, R. E., & de Croock, M. B. M. (2002). Blueprints for complex learning: The 4C/ID-model. *Educational Technology Research & Development*, 50(2), 39-64.
- Wideman, H. H., Owston, R. D., Brown, C., Kushniruk, A., Ho, F., & Pitts, K. C. (2007). Unpacking the potential of educational gaming: A new tool for gaming research. *Simulation & Gaming*, 38, 10-30.

- Wiliam, D., Lee, C., Harrison, C., & Black, P. (2004, March). Teachers developing assessment for learning: Impact on student achievement. *Assessment in Education, 11*(1), 49-65.
- Young, M. (1995). Assessment of situated learning using computer environments. *Journal of Science Education and Technology, 4*(1), 89-96.