Consequential Decisions from Continuously-Gathered Electronic Interactions: Could it Really Work?

Randy Elliot Bennett
Educational Testing Service
Princeton, NJ 08541
rbennett@ets.org

Paper presented at the annual meeting of the International Association for Educational Assessment, Singapore, May 2014.

Abstract

In an online learning environment, assessment information can be gathered continuously, ubiquitously, and unobtrusively. That information gathering can occur through an e-book, online course, game, or simulation. Some commentators have suggested that this capability will lead to the "end of testing." That is, there will be no reason to have distinct formative or summative assessments because all the information needed for classroom decision-making, as well as for student and institutional accountability, will be gathered in the learning process. If a student's continuously gathered electronic interactions suggest mastery of target competencies, why couldn't that information also suffice for promotion, graduation, college admissions, and teacher and school evaluation?

Whereas this idea seems compelling, there are at least five significant--and possibly intractable--issues that must be resolved before continuously embedded assessment can become reality in the summative context. These issues relate to the extrapolation of within-environment performance to outside performance, the comparability of performance across electronic learning environments, the privacy of the student data collected, the impact on teaching and learning, and the effect on the integrity of the formative and summative assessment processes. This paper will review these issues and pose one potential solution.

Keywords: technology based assessment, games, continuous assessment

The Common Core State Assessments (CCSA) represent the most significant change in US assessment at the primary and secondary school level in decades. One of the more obvious changes is their alignment with the Common Core State Standards (CCSSI, 2010), a uniform set of curriculum goals adopted by most states. A second change is the use of technology for test delivery, including computerized adaptive testing, for one of the two main state assessment consortia. Less noticed, is that the CCSA have done away with the traditional one-time test. Both main consortia, PARCC and Smarter Balanced, have divided their (long) tests into two administrations given at notably different points in time. In the case of PARCC, the administrations will occur after 75% and 90% of the school year (PARCC, 2013), with the parts aggregated to create a single proficiency estimate. Long tests and multiple sittings are necessary because, when the target of inference is the individual, a domain cannot be measured both broadly and deeply in any other fashion. This reality is at the root of the distributed testing regimes common in the assessment of individuals for professional licensure and certification (Bennett, in press).

Whereas the CCSA will be distributed over two time points, we now have the technology to sample student performance far more frequently, even continuously. Continuous assessment will be made possible by the emergence of electronic learning environments--e-books, online courses (e.g., massive open online courses, or MOOCs), simulations, and games--into which assessment can be embedded. In such environments, every event can be recorded, providing a massive quantity of information (Bennett, press). To the degree that students do all, or even a significant portion of their learning in such environments, the collected results (or log file) might constitute fine-grained documentation of accomplishment.

To what extent might that documentation add to--or even take the place of--formative and summative assessment as now conducted? For three decades, the field of intelligent tutoring has used the analysis of student responses in electronic learning environments to regulate instruction dynamically (Sleeman & Brown, 1982). The most successful of these intelligent systems in terms of both market share and evaluations of effectiveness are the Carnegie Learning Cognitive Tutors (Anderson et al., 1995; Pane et al., 2013; Ritter et al., 2007). These tutors, which were originally intended to be self-contained, are now meant to be employed in conjunction with teacher-directed activity. That change from teacher substitute to complement was likely motivated by the recognition that learning is a social activity best pursued through communities of practice (Lave, 1991; Wenger, 2000). Those communities are defined by their members--students, peers, and teachers--interacting around domain competencies, learning goals, valued tasks, student work, problem-solving approaches, and how that work and those approaches might be made better. The highly social nature of this interaction can certainly be enhanced by technology. For example, one could envision the log file being employed as input to the learning community (including the broadening of it beyond the school), and as input to the teacher's decision-making about how to regulate instruction (within and beyond the electronic environment). Even so, it is unlikely that technology can replace the community and the role that an experienced instructor has in overseeing it.

To be most effective, the use of log files presumes that the electronic learning environment be created from a carefully done domain analysis. In addition, the learning environment would need to use tasks devised to provide evidence of student standing on key domain competencies that came out of that analysis. Such theoretically motivated design should

increase the likelihood of our being able to make sense of the thousands of events that could potentially comprise such a file. Using empirical associations alone, as in data mining, may not generate results that are either meaningful or educationally reasonable (though it could lead to hypotheses worth testing). Finally, as with formative assessment generally, the inferences about student standing, and about the associated instructional adjustments, must be replicable and effective. In particular, greater learning should result from employing those judgments than from ignoring them (Bennett, 2011).

If the log file did prove to be valid for formative purposes, might it also be valid for more consequential purposes? Some investigators have proposed that analyses of student responses in learning environments might replace summative assessment (e.g., Bennett, 1998, pp. 11-14; Gee & Shaffer, 2010; Pellegrino, Chudowsky, & Glaser, 2001, pp. 283-287; Tucker, 2012). If a log file maintained for formative purposes incidentally produces evidence that a student is competent, wouldn't that demonstration eliminate the need for additional assessment? After all, the amount of responding related to both end result and problem-solving process would appear to make inferences about proficiency more dependable and meaningful. Whereas this notion is attractive, several issues would need to be addressed for it to become viable.

The first issue is extrapolation (Kane, 2006), or what accomplishment in a given electronic environment says about accomplishment *beyond* that environment. Meaningful extrapolation requires that the content targeted by a learning environment be closely mapped to relevant content standards like the CCSS. Extrapolation further requires that accomplishment in the environment is measured both validly and dependably. These requirements should by no means be taken for granted. Without evidence, what accomplishment in one learning environment says about the probability of accomplishment in the next grade, in college, or in a career is not known. Because of local control in US education, pupils in a state or consortium will interact with hundreds, if not thousands, of distinct simulations, games, online courses, or e-books. Validating the extrapolation inference empirically for the measures coming from each and every such learning environment will simply not be feasible.

Taking as given the relationship between accomplishment in electronic learning environments and achievement in criterion situations, a second challenge must still be met. This challenge concerns whether performance across electronic learning environments can be considered to be exchangeable. Even when created from the same set of content standards, like the CCSS, learning environments will vary noticeably in the distribution of standards and their depth of coverage, what tools and knowledge representations are employed, and the problem formats and contexts that are used. The questions included will inevitably be targeted at a certain collection of standards, tools, representations, formats, and contexts, making it hard to compare performance results for individuals working in different environments. A solution to this problem has existed for a long time. If the log file captures a pupil's work for the year, the summary score produced from performance in an electronic learning environment will be similar to a course grade. It is widely known that course grades vary considerably in meaning across teachers and schools (USDOE, 1994; Woodruff & Ziomek, 2004) and, consequently, can produce inequity when used for consequential decision-making. As a result, for such uses as postsecondary admissions, average course grades are typically scaled through a common measure (i.e., the SAT or ACT).

The third challenge can be stated as follows: Does the recording of essentially *every* learning and teaching behavior go against common expectations for privacy, particularly if used

for making decisions that affect the life-chances of individuals (Pellegrino, Chudowsky, & Glaser, 2001, p. 287)? Such monitoring would appear to require informed consent, which suggests that one may choose not to participate. Further, the public school would seem to be an instrument of the state and it is not immediately evident that the state has the authority to monitor one's cognition constantly. In the US, public behavior can generally be so observed but, without a court order, private behavior cannot be monitored. Should learning behavior, when continuously observed, be considered as public or private, especially if it is used for consequential decisions about individuals? Last, if student cognition can be constantly monitored, to whom should those data be made available? Clearly, most school districts will not have the resources required to manage and store those data. The rules for the federal *Family Educational Rights and Privacy Act*, which allow schools to outsource the management of such information without parental notification, have further heightened privacy concerns (Singer, 2013).

A fourth concern is one of effect on learning and teaching. Might the advantages of preparing for, and taking, a culminating examination be diluted if only continuously embedded assessment was employed? If the test accurately portrays content standards, studying for it should produce a significant beneficial result. Research shows that practice aids students in reaching proficiency--i.e., it helps them develop automaticity for basic skills, organize their knowledge, and link it to the situations in which it should be applied (Ericcson, Krampe, & Tesch-Romer, 1993; NRC, 2000).[1] Moreover, taking an examination can, itself, assist in the retention and transfer of learning (Butler, 2010; Roediger & Karpicke, 2006; Rohrer & Pashler, 2010).

The final concern relates to the idea that the most appropriate use of continuously embedded assessment is for guiding learning. Might employing formative results *incidentally* for consequential purposes compromise the efficacy of formative as well as summative decision-making? Effective instruction and learning require experimentation, or engagement in "productive failure" (Kapur, 2010). This useful activity could be suppressed by the awareness that all actions were being tabulated and evaluated. Worse, that awareness could potentially change the focus of formative assessment from enhancing learning and instruction to cynical attempts by students and teachers to subvert the system.

How might assessment programs address these issues as they evolve toward new forms of assessment? The "competitive-sports" metaphor is worth exploring as it employs continuous assessment, but with a clear differentiation between summative and formative uses. In baseball, for example, learning and practice--and the formative assessment connected to those activities--take place in spring training and, during regular season, between innings and between games. Formative assessment might also happen as a result of regulation play, but the overriding reason for that play is clearly summative--to decide what team wins the game, the championship titles, and the monetary rewards that come with those accomplishments. What matters in team performance, then, is assessed *only* during regulation play. The same situation is basically true for player performance.

[1] Practice certainly would occur in the context of continuous assessment. The concern expressed here is whether the elimination of the culminating examination would remove an additional, and highly beneficial, motivation to practice.

If the CCSA, for example, were to adopt the competitive-sports metaphor, it would employ continuous assessment in the electronic learning environments used in any given classroom, but only for formative purposes. Since schools would be free to buy *any* environment they desired, the embedded formative assessments would necessarily be provided by the learning-environment publishers. For summative purposes, however, the CCSA consortia might continue to offer their own examinations. Following the competitive-sports metaphor, there would be many examinations given throughout the school year, with the results aggregated within-student for purposes of portraying accomplishment (in the same way that teachers aggregate evidence to award an individual's course grade, or the major leagues aggregate wins and losses to determine which team goes to the Word Series). (See Mislevy and Zwick, 2012, for more on the psychometric challenges involved with such aggregation.) Also because of their number, each examination would have minimal impact, in contrast to today's one-time test, removing the problem of a student or class having a "bad day" (Bennett & Gitomer, 2009). These new assessments might also incidentally give tentative formative results, pointing toward a student's likely strengths and weaknesses, which the teacher would follow up with more targeted data gathering. The key distinctions between these CCSA-provided summative instruments and the publisher-embedded formative assessments would be that (1) the CCSA offerings would be common to all students regardless of what electronic learning environments they used and (2) students would know when their performance was being assessed for consequential purposes.

Summary and Conclusion

This paper critically reviewed the idea of using continuous assessment for summative, but also for formative, purposes. As suggested above, the challenges associated with using continuous assessment for consequential decision-making appear to be considerably more substantial than for classroom instruction. For consequential decision-making, the challenges relate to extrapolation of within-environment performance to outside performance, the comparability of performance across electronic learning environments, the privacy of the student data collected, the impact on teaching and learning, and the effect on the integrity of the formative and summative assessment processes. The "competitive sports" metaphor offers one possible approach to addressing these issues conceptually. Implementation, of course, would raise significant issues (e.g., how to aggregate results across assessments, the necessity for all schools to adhere to a given curricular sequence). These issues require much additional research and evaluation.

References

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *Journal of the Learning Sciences, 4*, 167-207.

Bennett, R. E. (1998). *Reinventing assessment: Speculations on the future of large-scale educational testing*. Princeton, NJ: Policy Information Center, Educational Testing Service. Retrieved January 21, 2014 from https://www.ets.org/research/policy_research_reports/pic-reinvent

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy and Practice 18*, 5-25.

Bennett, R. E. (in press). Preparing for the future: What educational assessment must do. *Teachers College Record*.

Bennett, R. E., & Gitomer, D. H. (2009). Transforming K-12 assessment: Integrating accountability testing, formative assessment, and professional support. In C. Wyatt-Smith & J. Cumming (Eds.), *Educational assessment in the 21st century* (pp. 43-61). New York: Springer.

Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36,* 1118–1133.

Common Core State Standards Initiative (CCSSI). (2010). *Common Core State Standards for English Language Arts and Literacy in History/Social Studies & Science.* Retrieved from http://www.corestandards.org/

Ericsson, K. A., Krampe, R. Th., & Tesch-Roemer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review, 100,* 363-406.

Gee, J. P., & Shaffer, D. W. (2010). Looking where the light is bad: Video games and the future of assessment. *Edge, 6(1),* 3-19. Retrieved January 16, 2014 from http://edgaps.org/gaps/wp-content/uploads/EDge-Light.pdf

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education/Praeger.

Kapur, M (2010). Productive failure in mathematical problem solving. Instructional Science, 38(6), 523-550.

Lave, J. (1991). Situating learning in communities of practice. In L. B. Resnick, J. M. Levine., & S. D. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 63-82). Washington, DC: American Psychological Association. doi: 10.1037/10096-003

Mislevy, R. J., & Zwick, R. (2012). Scaling linking, and reporting in a periodic assessment system. *Journal of Educational Measurement, 49,* 148-166.

National Research Council. (2000). *How People Learn: Brain, Mind, Experience, and School: Expanded Edition.* Washington, DC: The National Academies Press.

Pane, J. F., Griffin, B. A., McCaffrey, D. F., & Karam, R. (2013). *Effectiveness of Cognitive Tutor algebra I at scale* (WR-984-DEIES). Pittsburgh, PA: Rand Corporation. Retrieved February 19, 2014 from http://www.siia.net/visionk20/files/Effectiveness%20of%20Cognitive%20Tutor%20Algebra%20I%20at%20Scale.pdf

Partnership for Assessment of Readiness for College and Careers (PARCC). (2013). *PARCC assessment administration guidance (version 1.0)*. Author. Retrieved January 11, 2014 from http://www.parcconline.org/sites/parcc/files/PARCC%20Assessment%20Administration%20Guidance_FINAL_0.pdf

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: the science and design of educational assessment*. Washington, DC: National Academy Press.

Ritter, S., Anderson, J. R., Koedinger, K. R., & Corbett, A. (2007). Cognitive tutor: Applied research in mathematics education. *Psychonomic Bulletin and Review, 14,* 249-255

Roediger, H. L., III, & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17,* 249–255.

Rohrer, D., & Pashler, H. (2010). Recent research on human learning challenges conventional instructional strategies. *Educational Researcher, 39,* 406-412.

Singer, N. (2013, October 5). Deciding who sees student data. *New York Times*. Retrieved January 17, 2014 from http://www.nytimes.com/2013/10/06/business/deciding-who-sees-students-data.html?pagewanted=1&_r=0&adxnnlx=1389978450-gFo%20edDUCpRuRLvjV%20ngMQ

Tucker, B. (2012, May/June). Grand test auto: The end of testing. *Washington Monthly*. Retrieved January 15, 2014 from http://www.washingtonmonthly.com/magazine/mayjune_2012/special_report/grand_test_auto037192.php

United States Department of Education, Office of Educational Research and Improvement (1994). *What do students grades mean? Differences across schools*. (Office of Research Report OR 94-3401). Washington, DC: Office of Research. Retrieved January 17, 2014 from http://files.eric.ed.gov/fulltext/ED367666.pdf

Wenger, E. (2000). Communities of practice and social learning systems. *Organization, 7*, 225-246. Retrieved January 16, 2014 from http://org.sagepub.com/content/7/2/225 DOI: 10.1177/135050840072002

Woodruff, D. J., & Ziomek, R. L. (2004). *Differential grading standards among high schools* (ACT Research Report Series 2004-2). Iowa City, IA: ACT. Retrieved January 17, 2014 from http://www.act.org/research/researchers/reports/pdf/ACT_RR2004-2.pdf