Chapter 13
Epilogue: Achieving Quality 21st Century Assessment

Betsy Jane Becker and Valerie J. Shute

Writing a chapter to summarize all of the contents of this multifaceted book presents a challenge. Rather than create a laundry list of the conclusions drawn by our authors, we have tried to draw out three themes that appear across the works herein. These themes represent ideas that we believe will need careful attention from assessment experts, measurement professionals, teachers, principals, learning scientists and many others if the field is to move forward to develop better assessments that promote learning as well as provide fair means of accountability for students, teachers, and schools. We argue that

- Assessment must capitalize on advances in technology,
- Assessment is a contextualized, social activity, and
- Assessment must serve teaching and learning.

We discuss in turn how each of these themes is raised by the authors of our book, and also touch on potential areas for research suggested by their work. We do not, however, mention every instance in which every author touches on these themes. We apologize if we have omitted points on these themes that are important to our authors.

13.1 Assessment Must Capitalize on Technology

Technological advances have already clearly affected the world of assessment in many ways. Even mundane components of assessment, such as the scoring of multiple choice questions, were long ago made easier by Scantron machines and other scanning devices (Clarke, Madaus, Horn, & Ramos, 2000). However, while such devices enabled the rapid increase and widespread use of testing in the schools in the

B.J. Becker (✉)
Educational Psychology and Learning Systems Department, Florida State University, Tallahassee, FL, USA
e-mail: bbecker@fsu.edu

V.J. Shute, B.J. Becker (eds.), *Innovative Assessment for the 21st Century*,
DOI 10.1007/978-1-4419-6530-1_13,

1960s and 1970s, they did not always lead to improvements in what we know about students, or in what students learn (e.g., Epstein, Epstein, & Brosvic, 2001). Indeed, while modern technologies clearly have made an impact on how testing is done, it is clear they present both new challenges and new possibilities for assessment in the twenty-first century (e.g., Naglieri et al., 2004).

Hickey, Honeyford, Clinton and McWilliams examine a context where technology is inherently part of the assessment—the assessment of competence with new media. New media literacies include activities as diverse as social networking, creation of fan fiction, music remixing, and blogging. Because nearly all new media are based in technology, to assess competencies in these domains requires that technology be embedded in the assessment process. However, it is also true that traditional literacy skills—in writing, reading and spoken communication—are both needed for and enhanced by use of new media (see also Leu, O'Bryne, Zawilinski, McVerry, & Everett-Cacopardo, 2009). While endorsing the idea of assessing new media competencies, Hickey and his colleagues raise considerable concerns about whether the advent of accountability in this domain will narrow views of "proficiencies" to what is easily measured. Indeed, since many new media skills are inherently social (see our next theme), but most existing assessment systems are fundamentally individualized, clear tensions and conflicts will play out as assessment of these new skills moves forward. Given the view of many scholars that media literacy is (and must be) entwined with a participatory culture, such tensions will be a key concern for the field. But as with many challenges, also see interesting possibilities for research. For example, how should we best assess media literacy? Is it ever possible to gauge individual contributions to fully participatory activities? Many interesting research questions will emerge in relation to this context.

Several authors in our volume attend to the role technology must play in the future of assessment. An argument for an elegant system of evidence centered design (ECD) for assessment is given by Russell Almond. Almond lays out key aspects of the ECD philosophy he has developed with collaborators Robert Mislevy and Linda Steinberg. His chapter makes concrete how a complex mathematical modeling framework can be combined with thoughtful consideration of the skills desired of an examinee population to produce both improved learning and quality assessment all in one comprehensive system. The system relies on technology in its fundamental use of a Bayesian framework for evaluating student capacities. Information based on prior knowledge of examinee capabilities, along with data from observable events associated with a collection of tasks is fed back into a system to create posterior distributions of (hopefully changed) student skill levels. Almond also points out that an ECD system could aim at tracking growth in multiple competencies, based on related process or product observables, and can even interface with automated scoring systems (like those described by Shermis for essay scoring). Almond argues that eventually ECD, combined with modern technologies could support "seamless" (text page 30) collection of observables embedded in ongoing work—assessment that would seem so natural students would not realize that it had even occurred (see also Shute, in press for more on this topic).

Clearly such assessment would not cause the stress and disruptions present in so much high-stakes testing (e.g., Cizek & Burg, 2005; Suen & Yu, 2006).

While they endorse many of Almond's ideas, commenters Ellington and Verges point out that currently schools do not have the infrastructure to support these innovations. For example, problems and incompatibilities in hardware and software could lead to glitches with data collection from diverse sources. Also while they acknowledge the benefits of using complex tasks in assessments, they raise several practical problems (such as cost and extensive field test requirements) that would limit implementation especially in the current economic climate. Clearly such a system for assessment does not currently exist, but the goal of realizing it presents another set of fascinating possibilities for research.

A more conventional take on technology and assessment comes from Mark Shermis. His chapter describes how technology can assist with improved student learning of writing skills when the electronic scoring of essays is built into a system of writing improvement, along with revision, feedback, and teacher participation. Shermis describes how automated essay scoring or AES works, from the development of "proxes" or features used to represent the quality of a writing sample, to the evaluation of the statistical models used in algorithmic scoring. Shermis cites evidence that with careful development and good rubrics, an AES system is at least as reliable as human raters, and can in some cases help avoid biases that human raters cannot seem to eliminate from their rating behavior. One goal of this chapter is to simply describe how AES works, and Shermis further illustrates that by way of a detailed description of the "Intellimetric" system. However Shermis goes farther, and makes the controversial claim that automated essay scoring (and the teaching structures associated with its ongoing use) can replace high-stakes writing assessments. Consistent with Almond's arguments for incorporating multiple pieces of evidence in on going assessment system, Shermis argues that an assessment based on multiple instances of writing (such as essays produced throughout the year) would provide a more useful and valid evaluation of student writing than a single-occasion high stakes test. He describes how an integrated writing assessment system could also support instruction by providing feedback aimed at each essay produced by a student. In addition, the release of "used" writing prompts could provide materials for ongoing instruction. Finally he argues that all of this can be accomplished at costs lower than those incurred with human scoring of similar writing products. This is an excellent example of how capitalizing on technology can enhance not only the assessment itself but also student learning. And as with our other chapters that touch on technology, various ideas for research arise from Shermis's work. For example, what measurement models best suit this kind of assessment system? How often would one need to take samples of writing throughout the year to get solid information about change in writing competence? Could an AES system have a built-in way to assess the impact of particular kinds or amounts of teacher (or system) feedback? How would we assess whether the system is providing appropriate feedback to students (i.e., what kinds of "quality control" would be needed)? These and other practical and theoretical questions provide a rich set of ideas for those interested in the future of assessment.

13.2 Assessment is a Contextualized, Social Activity

The strongest advocate of the social view of assessment in our collection of authors is James Gee. Gee lays out the case for assessment of twenty-first century skills in domains as part of social "appreciative systems". Loosely these are sets of conventions, values—perhaps even rules—for what is acceptable or valuable in a certain domain. He argues that appreciative systems are shared across people, and gives quite a few examples of how such systems develop. He also argues that most assessment goes on as a part of ongoing human interactions and activity, and is not—and does not need to be—formalized. King Beach, who comments on Gee's chapter, provides several examples of such real-life assessment in out of school settings in western Nepal.

Gee also contends that groups themselves can formalize assessment in a quite natural way. He argues that this often occurs via "Pro-Am" communities, or groups of " ... innovative, committed and networked amateurs working to professional standards" (Leadbeater & Miller, 2004, p. 9). A compelling example comes from his research on online communities (Gee & Hayes, in press) where a young girl learned to create virtual clothing for the virtual world *Second Life*. Eventually various discoveries lead her to provide clothes to virtual people on the Internet, first for free then later at a price, after she realized that the "appreciative" community of virtual shoppers highly valued her product. More controversially, Gee goes on to argue that schools could promote twenty-first century skills by encouraging and equipping students to become high-status members of Pro-Am communities that value and promote such skills.

Gee notes that society has in some cases formalized assessment by removing it from such Pro-Am communities. Institutions, including schools, have been created in support of this formalization. But Gee believes this kind of assessment is "backwards" (text page 28) because it occurs at such an abstract, disembedded level, removed from real problems. His arguments for authentic assessments echo in part those of measurement scholars (e.g., Wiggins, 1990) and others (e.g., Darling-Hammond & Snyder, 2000) who called for more realism and context in assessment years ago.

Finally, Gee raises the radical idea that formalized assessments may not be needed if existing communities (like the online buyers of simulated clothing) have already assessed and accepted someone's skills. He states that, "The job of educators ought to be designing such social organizations and letting them run." (text page 32) Again we see plenty of possible research avenues in this work, and Beach's commentary raises one interesting question—what is the appropriate unit of assessment in such circumstances? Could we find a way to examine the developmental relationship between a learner and the domain to be assessed, over time? Could such a complex entity be assessed by the "indigenous workings" (Gee page 32) of the group, as advocated by Gee? It is intriguing to consider how one could study and obtain empirical evidence on such a system of assessment.

Hickey and colleagues also deal with the social context for learning, which is fundamentally a part of examining competencies with "new media". They begin by

examining the positions of groups like the National Council of Teachers of English and the National Writing Project, which argue that writing is an inherently social activity. This is a launching point for the efforts of Hickey and his colleagues to create a cohesively designed assessment framework for language arts. Their efforts have used not only a classic text (Moby Dick) but also new media sources such as related music videos and theatre "re-mixes". They argue that by using multiple levels of assessment (designed after Ruiz-Primo et al. 2002 five levels) the teacher can focus initially and closely on social and interactive aspects of writing, with a quick time frame for feedback and a very informal assessment context. The assessments and classroom activities then move to more individualized, but distant aspects at other levels. Their chapter gives many examples of assessment activities at their five levels, but the contextualized nature of assessment is most evident at their first two levels. Only at the highest level would fully individualized assessments (e.g., external tests with essay items) be used to tap into an abstract context, and might, for instance, measure performance on content standards such as those at a state level.

We consider one example from their second level—"close-level activity-oriented reflections" (page 29 of text). Here the focus is on discussion questions presented to the class, either orally, in written form, or online. The authors found that the informal nature of these assessments and an attendant emphasis on "communal discussion" of ongoing classroom activities enhanced broader student participation. The teacher then could consider the nature of the discussion, and focus further discussion in ways that led students ". . .to create more compelling and creative artifacts" (page 30 of text). This is an interesting example of how a teacher can assess student understanding in a group context, provide feedback based on questions that are less formal, but very targeted to the activities of the learners.

Allan Jeong's work in examining conceptual causal maps provides a nice example of how learning can be enhanced by the use of social interactions and by assessments that provide a window on the understandings of experts and of other learners. Jeong describes the use of jMAP, a program that enables learners to create and evaluate causal maps, and also allows for the assessment of changes in maps over time. Students use jMAP to identify critical components of a causal system, to draw interconnections among those components, and to identify the strength of those connections. The connection to "others" is embedded in the structure and use of jMAP. Specifically, after creating their own maps, students can be exposed to the maps of experts, to discussions about the nature of the causal connections (as was the case in Jeong's studies), and to composite maps made by aggregating the maps of various subsets of learners or all learners in a class. Subsequent rounds of maps can be drawn and changes in the maps of learners can be examined. Jeong shows how comparisons of individual versus aggregate maps can lead to changes in subsequent maps of individual learners. Also the nature of discussions held (e.g., whether links in the maps are supported or challenged by others, whether explanations are provided, etc.) can impact how learners change their maps on subsequent drawings.

For those interested in how learner interactions can impact learning, jMAP provides intriguing tools for analysis of causal maps. We challenge those interested in such phenomena to consider how learning in other domains of understanding (i.e.,

other than causal maps) might be represented (using technology) and then examined for changes due to exposure to the knowledge and ideas of others. Domains suggested by the work of our other authors might include writing, where automated scoring systems like those described by Shermis could produce indices of change following peer or teacher feedback of different sorts, or Gee's "Pro-Am skills" where changes in the products or skill sets of members of Pro-Am communities might be evaluated for evolution as individuals receive feedback from relevant community members. Roehrig and Christesen's Classroom AIMS is another assessment device where modal (typical) or expert performance ratings could be shared with teachers, and then further measurement instances could be evaluated for change due to those different kinds of feedback.

13.3 Assessment Must Serve Teaching and Learning

Mari Pearlman views assessment as integral to the educational enterprise. In her chapter she argues that we must "...ally assessment with instruction" (page 20 of text version) and she makes a provocative case for using an assessment based on the vast architecture of the National Assessment of Educational Progress or NAEP as a vehicle to accomplish this goal. Pearlman outlines a variety of problems in our educational system, and among them she lists a need to align all components of the system—curriculum, teacher practice, teacher preparation and assessments—in concert to move student learning ahead. She argues that to date these components have been manipulated by the states, but mainly in efforts to achieve "AYP" or adequate yearly progress, not in efforts to increase learning of content identified as important (See for example Kane, Staiger, & Geppert's, 2002 views on gaming the AYP system so states can look most successful). Pearlman endorses the move towards national standards, and argues that having clear frameworks could lead to national "conversations" about content, curriculum and most importantly equity across states. Equity is an issue conveniently (but sadly) avoided in the current context, where each state can use its own distinct tests to assess progress, and also can set different goals. In such a context, it is not hard to see that all states could theoretically measure up as "adequate" while in fact being quite different in terms of what their students actually learn. Initial examinations however showed that some states did the opposite—setting such high standards that they were virtually unreachable (Linn, Baker, & Betebenner, 2002).

In terms of the theme of assessment in support of learning, Pearlman challenges us to find ways to use the many items developed as part of the NAEP assessments in support of learning. Indeed, thousands of released NAEP items are currently available for public use via the NAEP Questions Tool (see http://nces.ed.gov/nationsreportcard/itmrlsx/landing.aspx), and can be used by teachers and others in a variety of ways. Much like the position taken by Almond, Pearlman argues that we must first know what we want students to learn, then what we want to measure, then only last can we design lessons and activities to support

those goals. Pearlman argues that if benchmarks such as those set for NAEP were used as goals, and NAEP's current plan of sampling students were replaced with every-pupil-testing, we would soon move towards better outcomes and towards equity across states in terms of student learning.

In her response to Pearlman, Lynn Wicker emphasizes a point made by Pearlman. That is, both agree that the culture of schools relative to assessment must change so that a stronger link to learning can be achieved. Wicker argues that the use of multiple ongoing assessments tied with targeted interventions "should be viewed as a non-negotiable in the learning process" (page 6 of text version). We also strongly endorse this view.

Martineau and Dean give perhaps the most comprehensive proposal in our volume for how assessment, in many forms and at many levels, can serve instruction and learning. They draw on the idea of balanced assessment (Redfield, Roeber, Stiggins, & Philip, 2008) and elaborate it to describe how formative, summative and interim assessments can be used to provide both indices for accountability as well as detailed input to teachers in support of their instruction. Their system begins with clear and focused K-12 content standards, aimed at supporting students' progression towards specified high-school outcomes. They argue that curriculum materials ("model curriculum units") can then be developed in support of these standards, and be made available to all teachers (but not mandated as a required curriculum). This set of content standards and materials would be paired with professional development for teachers aimed at helping them to understand the content standards and how they can be used, but more critically how to use data from assessments—both classroom assessments and more formalized "secure" assessments to modify their instruction.

A requirement that teachers receive instruction in classroom assessment is already a part of the Florida Department of Education's preservice teaching requirements. Martineau and Dean want continued support in the form of professional development for both teachers and other administrators. They also argue that accountability purposes can also be served by their system, but only with a multifaceted system in which teachers, teacher preparation institutions, administrators and students are all held to account.

Working from a very different perspective, Hickey and colleagues also describe a system of assessment that focuses on different kinds and levels of assessment in the context of new media literacy. Based in the technological context discussed earlier, the first three of their five levels of assessment—the immediate, close and proximal levels—are completely entwined with student products (artifacts) and interactions with teachers and other students as they create those products in the classroom. At the third (proximal) level the assessment tasks move towards more individualization, but even at this level reflection questions and student products together allow for targeted teacher feedback. Their system illustrates how assessment can be tied directly into ongoing instruction.

A more targeted approach to both assessing and supporting instruction is described by Roehrig and Christesen. They describe the development and use of the Classroom AIMS instrument, which examines how teachers create a

positive classroom atmosphere, implement instruction and classroom management, and engage students in learning. Roehrig and Christesen start with the premise that teachers' behaviors and activities are more likely to predict student outcomes than teacher characteristics such as teacher knowledge. They describe how a set of exemplary behaviors were identified by observing teachers who had produced students with strong learning gains in reading and writing performance. More importantly for this "theme" of our book, the authors go on to describe how the AIMS instrument can be used not just to observe the current behaviors of teachers, but also to diagnose possible areas for teachers to improve. They describe the use of the AIMS with preservice teachers as well as practicing teachers working with mentors. Other research questions could be asked about the Classroom AIMS. How does AIMS function as a measure of the effects of professional development for practicing teachers? Can it differentiate between more and less effective interventions? How well does it work across different subject areas? These and other questions may be fertile areas for future research.

13.4 Conclusion

As mentioned in the Prelude, our goal for the symposium and this book was to bring together groups of individuals who normally do not converse, but who we believe should communicate—researchers from different areas, policymakers, and educational professionals. In all chapters, the call for educational reform is clear and there is no shortage of problems to be addressed with innovative thinking and high-quality research. The linchpin for such reform—reform that aims to fully support students' success in the twenty-first Century—is assessment.

The chapters in this book present a broad swath of assessment issues and possible solutions, and embrace three main theses: (a) assessment must capitalize on advances in technology, (b) assessment is a contextualized, social activity, and (c) assessment must serve teaching and learning. Each of these alone can move the assessment conversation and ensuing research forward, but we contend that when these issues are considered collectively, important breakthroughs in assessment and educational reform will surely follow.

References

Cizek, G. J., & Burg, S. S. (2005). *Addressing test anxiety in a high-stakes environment: Strategies for classrooms and schools*. Thousand Oaks, CA: Corwin Press.

Clarke, M. M., Madaus, G. F., Horn, C. L., & Ramos, M. A. (2000). Retrospective on educational testing and assessment in the 20th century. *Journal of Curriculum Studies*, *32*(2), 159–181.

Darling-Hammond, L., & Snyder, J. (2000). Authentic assessment of teaching in context. *Teaching and Teacher Education*, *16*(5–6), 523–545.

Epstein, M. L., Epstein, B. B., & Brosvic, G. M. (2001). Immediate feedback during academic testing. *Psychological Reports*, *88*(3), 889–894.

Kane, T. J., Staiger, D. O., & Geppert, J. (2002). Randomly accountable. *Education Next*, *2*(1), Downloaded 12/16/09 from http://educationnext.org/randomly-accountable/

Leadbeater, C., & Miller, P. (2004). *The Pro-Am Revolution: How enthusiasts are changing our economy and society*. Demos. Downloaded on January 12, 2010 from http://www.demos.co.uk/files/proamrevolutionfinal.pdf?1240939425

Leu, D. J., O'Bryne, I., Zawilinski, J., McVerry, J. G., & Everett-Cacopardo, H. (2009). Expanding the new literacies conversation. *Educational Researcher*, *38*(4), 264–269.

Linn, R. L., Baker, E. L., & Betebenner, D. W. (2002). Accountability systems: Implications of requirements of the No Child Left Behind Act of 2001. *Educational Researcher*, *31*(6), 3–16.

Naglieri, J. A., Drasgow, F., Schmit, M., Handler, L., Prifitera, A., Margolis, A., et al. (2004). Psychological testing on the Internet: New problems, old issues. *American Psychologist*, *59*(3), 150–162.

Redfield, D., Roeber, E., Stiggins, R., & Philip, F. (2008). *Building balanced assessment systems to guide educational improvement*. A background paper for the keynote panel presentation at the National Conference on Student Assessment, June 15, 2008, Orlando, FL. Downloaded on December 15, 2009, from http://www.ccsso.org/content/PDFs/OpeningSessionPaper-Final.pdf

Ruiz-Primo, M. A., Shavelson, R. J., Hamilton, L., & Klein, S. (2002). On the evaluation of systemic science education reform: Searching for instructional sensitivity. *Journal of Research in Science Teaching*, *39*(5), 369–393.

Suen, H. K., & Yu, L. (2006). Chronic consequences of high-stakes testing? Lessons from the Chinese Civil Service Exam. *Comparative Education Review*, *50*(1), 46–65.

Wiggins, G. (1990). *The name assigned to the document by the author. This field may also contain sub-titles, series names, and report numbers. The case for authentic assessment*. ERIC Digest. ERIC Document Reproduction Service Number ED328611. Washington, DC: ERIC Clearinghouse on Tests Measurement and Evaluation.

Gee, J. P., & Hayes, E. R. (in press). *Women and gaming: The Sims and 21st century learning*. New York: Palgrave/McMillan.

Shute, V. J. (in press). Stealth assessment in computer-based games to support learning. To appear In S. Tobias & J. D. Fletcher (Eds.), *Computer games and instruction*. Charlotte, NC: Information Age Publishers.

AQ3

AQ4

Chapter 13

Q. No.	Query
AQ1	There is mismatch in the chapter title between the chapter opening page and contents. We have followed as per Chapter opening page. Please confirm
AQ2	Please provide Abstract and Keywords for this chapter in order to maintain book consistency
AQ3	Please update the year for the reference “Gee and Hayes, in press”
AQ4	Please update the year for the reference “Shute, in press”