Center for
K–12 Assessment
& Performance Management

An independent catalyst and resource for the improvement of measurement and data systems to enhance student achievement.

Exploratory Seminar:

Measurement Challenges Within the Race to the Top Agenda

December 2009

Recommendations for High-Quality Instructional Guidance Assessment Systems and How They Might Articulate With an Accountability System: Discussion

Randy E. Bennett

Created by Educational Testing Service (ETS) to forward a larger social mission, the Center for K–12 Assessment & Performance Management has been given the directive to serve as an independent catalyst and resource for the improvement of measurement and data systems to enhance student achievement.

Recommendations for High-Quality Instructional Guidance Assessment Systems and How They Might Articulate With an Accountability System: Discussion

Randy E. Bennett

Educational Testing Service, Princeton, New Jersey

This paper is based on a reaction by Randy E. Bennett to presentations by Margaret Heritage, Lauren Resnick, and Mark Wilson at the Exploratory Seminar: Measurement Challenges Within the Race to the Top Agenda, December 2009. Download copies of the papers presented at the seminar at http://www.k12center.org/publications.html.

My reaction to the three interesting and provocative presentations by Margaret Heritage, Lauren Resnick, and Mark Wilson takes the form of a metaprinciple followed by eight consequent principles. The metaprinciple and the consequent principles are derived from work we have been undertaking to create new approaches to K–12 assessment (Bennett & Gitomer, 2009).

The metaprinciple is itself a principle, but one so important that I choose to elevate it because it frames what follows. The metaprinciple is that *summative, interim, and formative assessment should be designed as complementary components of the same coherent system*. The reasoning behind this metaprinciple is based on the claim that formative and interim assessment will inevitably gravitate toward the design, format, and content of the summative system, and that without significant attention to the design of summative assessment, formative and interim assessments' potential to positively affect learning and instruction will be severely limited.

The implication of this metaprinciple is that one needs to set design constraints on the summative system so that it works with, and *not* against, the system components directly concerned with instructional guidance. Stated another way, one cannot talk meaningfully about "instructional guidance assessment systems" without considering the summative component. Each of the presenters articulated this principle in one form or another. For example, Heritage offered an integrated framework for three levels of assessment, Wilson spoke of building assessment from classroom curriculum upward, and Resnick implied coherence in terms of the relationship of assessment to curriculum and instruction.

In keeping with the coherence metaprinciple, my first two principles focus on facilitating synergy among assessment system components. Principle 1 is that *summative, interim, and formative assessment should be built from the same (strong) conceptual base.* That base should consist not only of content standards or curriculum (as in Resnick's formulation), but also of results from cognitive-scientific research. That research offers cognitive-domain models which postulate how the components of domain proficiency—processes, strategies, knowledge, and habits of mind—might be organized, and how those domain components might work together to facilitate skilled performance. That research also

offers learning progressions or developmental models that suggest how domain components might change over time.

All three presenters invoked the notion of learning progressions, Heritage describing it as a "connected network that represents how competence in a domain develops" and Wilson as a "description of the successively more sophisticated ways of thinking about an idea that follow one another as students learn." Wilson saw learning progressions as providing the potential to improve the match between curriculum and assessment, address the "mile-wide/inch-deep curriculum problem," increase assessment efficiency, improve teacher interpretability of assessment results, and allow a long-term view of student growth. Of note is that Wilson's conception, as least in its examples, implied a relatively fine-grained, within-grade progression keyed to a core understanding or big idea. Resnick's conception, in contrast, appeared to be more global and more concerned with articulating the development in domain competence across the K–12 range of schooling.

To be sure, learning progressions are a very promising approach, but there appear to be fewer empirically supported examples than there are core topics in any given curriculum, so it might be best to view learning progressions as an emerging idea around which much work still needs to be done.

I want to give extended attention to my second principle because I think it is so critical and because I believe that the measurement field has so far failed to address it effectively. This principle has several parts, which each of our presenters addressed, though perhaps from a direction slightly different than the one I take.

The principle is that summative and interim assessment should be designed to support instruction to the maximum degree possible (without compromising their primary purposes).[1] The first part of this principle is that these components should be worthwhile educational experiences. As an example, students should learn by preparing for these assessments. Such preparation will be most beneficial, of course, to the extent that the assessments are a reasonably rich representation of the standards or curriculum they are intended to measure. Resnick described such preparation as a good version of teaching to the test. I would suggest that, rather than encouraging teaching to the test, we would do better to encourage "teaching to the construct" (Messick, 1992, pp. 9, 16). That is, we want teachers and students to focus upon the competencies represented by the test, not upon the tasks or task formats themselves.

In addition to learning through preparation, students should learn from taking those assessments, because summative and formative assessments should give students something nontrivial with which to reason, think critically, write, or do mathematics. Figure 1 gives an example from an ETS research initiative called *Cognitively Based Assessment of, for, and as Learning* (CBAL; Bennett & Gitomer, 2009).

The screenshot in Figure 1 summarizes a portion of a prototype middle-school English language arts assessment containing a sequence of tasks focused on a single topic, e-waste. Among other things, the student needs to listen to an online radio news report about e-waste and take notes on its content,

[1] A primary purpose of interim assessment is to predict which students are in danger of failing to meet proficiency so that remedial resources can be brought to bear.

Figure 1. An example scenario-based task set illustrating the range of materials with which students must engage for assessment purposes. The task set depicted in this screenshot is from Sheehan and O'Reilly (2008).

evaluate websites that contain information about e-waste, read articles and watch a video about e-waste and then answer questions about them, use a graphic organizer to help manage information from these different sources, and write a well-informed letter to a computer company with recommendations for programs that the company might adopt to help address the e-waste problem. The point of this example is that the content students encounter is extensive and nontrivial, so that by the time the student is done with this assessment, he or she is likely to know a lot more about the e-waste issue than before.

The second part of the support instruction principle is that *summative and interim assessment should model good instructional and learning practice*. Such modeling can be achieved by including tools and representations that proficient performers work with, and by encouraging the habits of mind common to proficient performers in the domain. Figure 2 gives an example of an assessment task in which we model the use of a tool for analyzing complex text by asking students to complete a very simple graphic organizer. (The reference text is not shown in Figure 2, but it can be found on the tab labeled *Drowning in E-waste*.) In this task the student is given superordinate categories and must fill in details based on the reference text. In other tasks the student is given details and must fill in the superordinates.

Figure 3 shows how students are encouraged to use the planning tool of their choice for purposes of organizing their thoughts in preparation for writing. They can choose from among List, Free Writing, Idea Tree, Idea Web, and Outline tools.

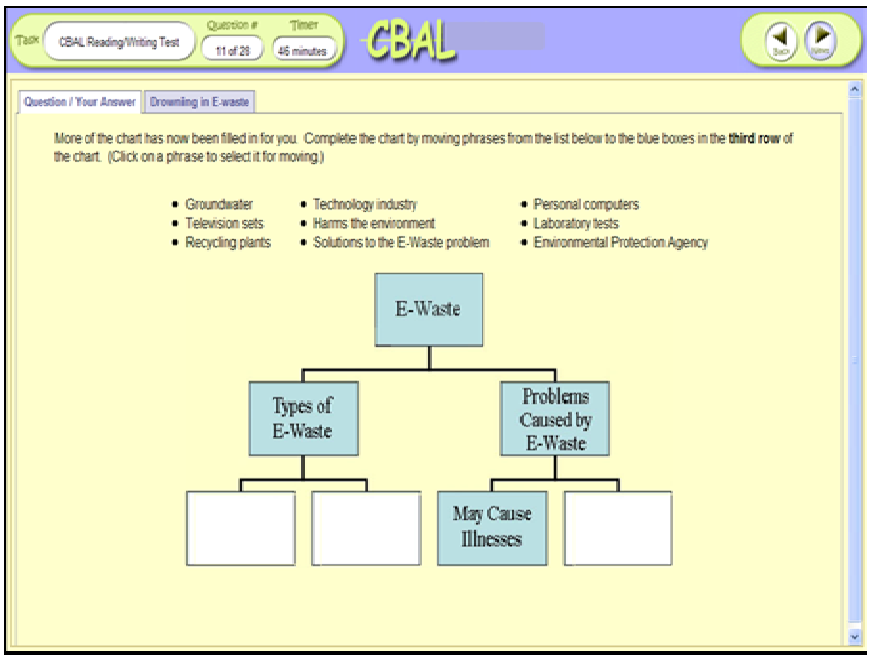

Figure 2. An assessment task that models the use of a graphic organizer for analyzing complex text. The task set depicted in this screenshot is from Sheehan, O'Reilly, Nadelman, Elkins, and Fowles (2009).

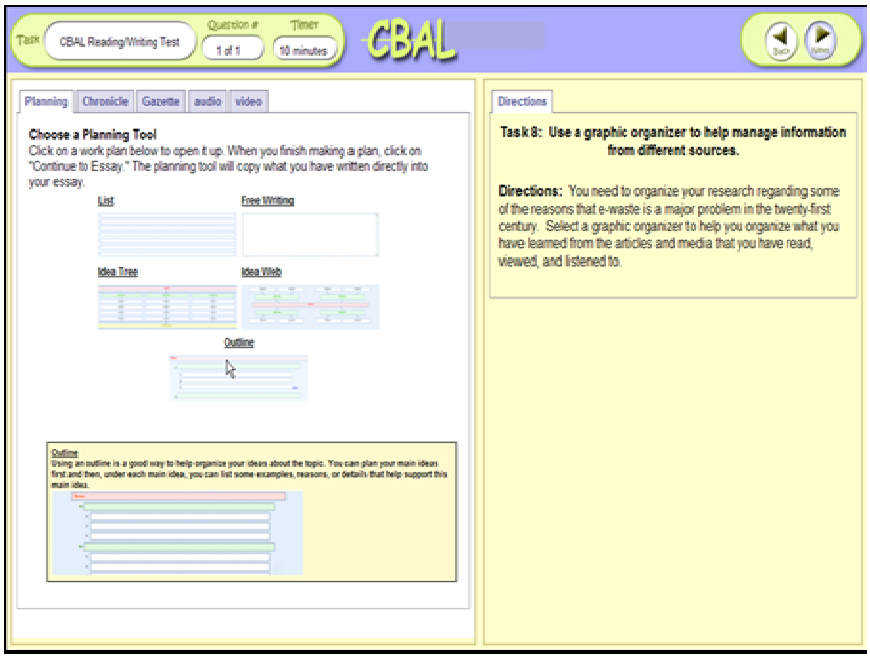

Figure 3. An assessment task that models the use of planning tools for writing. The task set depicted in this screenshot was created by P. Deane, M. Fowles, and J. Burstein of ETS.

On the left of the screen presented in Figure 4, students are given a set of criteria for evaluating the quality of Internet sources, which they are then asked to apply to website summaries, shown on the right. The hope is that repeated exposure to such criteria will help students internalize them so that they become a habit of mind.

The last component of the support instruction principle is that summative and interim assessment should tentatively point teachers (and students) to areas of potential concern for more targeted follow-up. This principle is similar, I think, to what Linn and Betebenner in their presentation called descriptive use in the context of growth modeling. In the context of summative and interim assessment I call these pointers formative hypotheses that teachers and students can pursue through classroom assessment (Bennett, 2009).

Lorrie Shepard (2006) touched on this idea in her "Classroom Assessment" chapter from the fourth edition of *Educational Measurement.* She wrote:

> I see a strong connection between . . . formative assessment practices . . . and my training as a clinician when I used observations to form a tentative hypothesis, gathered additional information to confirm or revise, and planned an intervention (itself a working hypothesis). (p. 642)

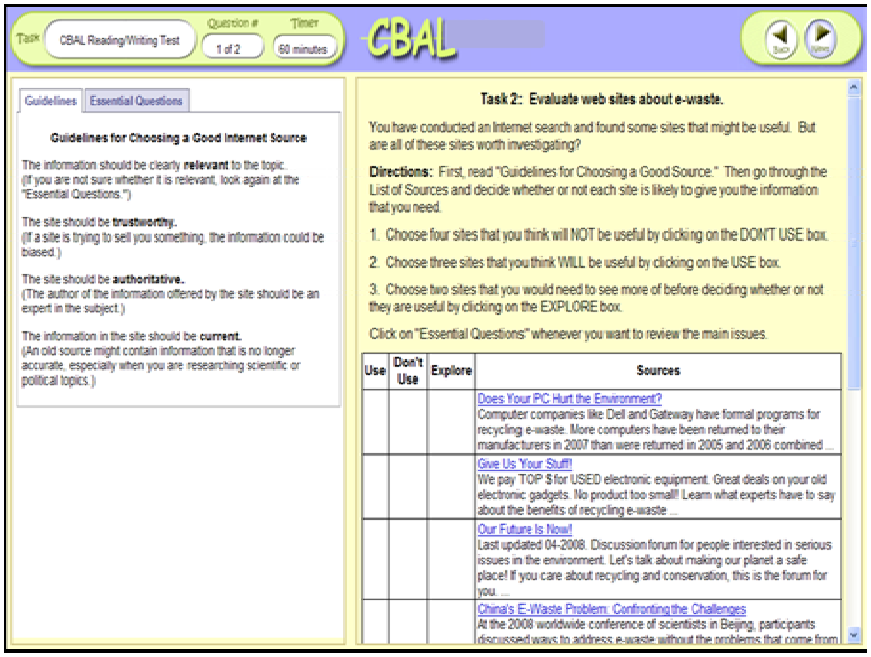

Figure 4. An assessment task modeling the application of criteria for evaluating the quality of work. The task set depicted in this screenshot is from Sheehan et al. (2009).

The importance of such hypotheses becomes clearer when we think about the distinctions among *errors*, *slips*, *misconceptions*, and *lack of understanding*. An error is what we observe students make, some discrepancy between an expected response and what the student gives. A summative or interim assessment can easily identify the errors a student makes and, perhaps also, some of the content classes in which those errors appear to occur disproportionately.

Any error we observe a student make, however, may have one of several causes. Among other things, that error could be the result of a slip—that is, a careless procedural mistake. It could be due to a misconception—some persistent conceptual or procedural confusion. Or it could be caused by a lack of understanding—a missing piece of conceptual or procedural knowledge, without any persistent confusion. Each of these causes implies a very different instructional action, from minimal feedback (for the slip), to re-teaching (for the lack of understanding), to the significant investment required to engineer a deeper cognitive shift (for the misconception). What a summative or interim assessment is very unlikely to be able to do is identify such causes.

The key point is that carefully designed summative and interim assessments ought to be able to generate initial formative hypotheses (e.g., about content classes that maybe problematic for individuals or groups). The teacher's role is to confirm or refute those initial hypotheses, and to attempt to isolate the causes, through classroom formative assessment. That formative assessment could entail asking for the student's explanation as to why he or she chose to respond in a particular way, administering more tasks and looking for a pattern of responses consistent with the hypothesis, and/or relating the error to other examples from the student's class work, homework, or test performance.

Will this be easy for the typical teacher to do effectively, even at the group level? Probably not, but it is time that preservice training institutions, testing companies, state departments, and the federal government made a concerted, collaborative attempt to help teachers develop these critical competencies.

Principle 3 is that *formative assessment should use a range of task types and modes (as appropriate to the competencies required for domain proficiency).* Task types might include multiple-choice and short constructed-response items; extended, scenario-based tasks; projects; and portfolios. Modes might include individual as well as collaborative and non-collaborative group work, and interactive discussion as well as seat work.

Principle 4 states that *formative assessment should use technology* (*where technology can make a meaningful contribution).* For example, technology can make possible the measurement of procedural fluency and the measurement of problem-solving processes. It can give students the opportunity to practice on constructed-response tasks and provide automated feedback or, going a step further, through "intelligent tutoring" it can select instructional exercises keyed to the student's particular constellation of knowledge and skill. Technology can allow students to score benchmark constructed responses, the intention here being to help students learn to recognize good work and its characteristics. Finally, it can allow students to create (and reflect upon) digital portfolios.

Only Heritage directly addressed this principle, discussing the use of technology to administer dynamic assessments and the digital collection of artifacts (including audio and video) for accountability and for the consolidation of learning.

Principle 5 deals with an issue that I would like to have heard the presenters directly discuss. The principle is that *formative assessment should not be used for accountability purposes*. My reasons are that fairness would seem to demand that teachers and students know when they are being assessed for consequential (as opposed to learning) purposes. I also believe that if the purpose of formative assessment changes to include the consequential uses of results, student and teacher behavior will be affected in ways that reduce the original, instructional intent of the practice. As a simple example, students and teachers need room to experiment in learning, including trying out approaches that may fail. That type of experimentation, I would argue, is less likely to occur if formative assessment becomes an accountability component.

Principle 6 is that *teachers (and students) should have maximum flexibility in the use of formative assessment*. Researchers and vendors should provide infrastructure for implementing, adapting, and sharing formative assessment materials, and for managing results, including conceptual models, tools, tasks, and data systems. Whether, how, and when to use that infrastructure should be a local (preferably classroom) decision, however. My reason for this principle is that as good as our models, tools, tasks, and data systems turn out to be, there will be teachers with ideas better suited to their particular classroom circumstances. Although I believe it was intended by each of the speakers, there was not much direct reference to this "flexibility" principle in the presentations.

Principle 7 is that *by design, formative assessment should encourage students to take part in formative assessment*. A commonly articulated goal is that formative assessment should help students to internalize standards and criteria for proficient performance, and to routinely use them in evaluating their and their peers' work. Heritage gave direct attention to this principle with the statement that students should be involved with their teachers in assessing and monitoring their learning, especially in Levels 2 and 3 of her integrated framework, where students and teachers appear to work as partners in assessment.

My last principle states that *summative, interim, and formative assessment should be development experiences for teachers*. That is, these assessment opportunities should focus attention on, and provide examples of, key standards, cognitive-domain models, and learning progressions; this would encourage the development of a deeper understanding of the domain being taught and, hopefully, also encourage development of assessment fundamentals. I would like to have heard more attention given to this principle in the presentations.

In summary, my recommendations for high-quality instructional guidance assessments systems are that summative, interim, and formative assessment should be designed as complementary components of the same coherent system, and that these components should be built from the same (strong) conceptual base. Summative and interim assessment should be designed to support instruction. Formative assessment should use a range of task types and modes; use technology where technology can make a meaningful contribution; *not* be used for accountability purposes; allow teachers (and

students) maximum flexibility in its use; and, by design, encourage students to take part in it. Finally, summative, interim, and formative assessment should be development experiences for teachers.

References

Bennett, R. E. (2009). *A critical look at the meaning and basis of formative assessment* (ETS Research Memorandum No. RM-09-06). Princeton, NJ: ETS.

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, NY: Springer.

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

Messick, S. (1992). The interplay of evidence and consequences in the validation of performance assessments (RR-92-39). Princeton, NJ: ETS.

Sheehan, K. M., & O'Reilly, T. (2008, April). *The case for scenario-based assessments of reading competency*. Paper presented at the Assessing Reading in the 21st Century Conference, Philadelphia, PA.

Sheehan, K. M., O'Reilly, T., Nadelman, H., Elkins, B., & Fowles, M. (2009, June). *The CBAL Reading Assessment: An approach for balancing measurement and learning goals.* National Conference on Student Assessment, Los Angeles, CA.

Shepard, L. A. (2006). Classroom assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 623–646). Westport, CT: Praeger.