CRESST REPORT 739

Terry P. Vendlinski
David Niemi
Jia Wang
Sara Monempour

IMPROVING FORMATIVE ASSESSMENT PRACTICE WITH EDUCATIONAL INFORMATION TECHNOLOGY

JULY, 2008

National Center for Research on Evaluation, Standards, and Student Testing

Graduate School of Education & Information Studies
UCLA | University of California, Los Angeles

Improving Formative Assessment Practice with Educational Information Technology

CRESST Report 739

Terry P. Vendlinski, David Niemi, Jia Wang, & Sara Monempour
CRESST/University of California, Los Angeles

July 2008

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Center for the Study of Evaluation (CSE)
Graduate School of Education & Information Studies
University of California, Los Angeles
300 Charles E. Young Drive North
GSE&IS Bldg., Box 951522
Los Angeles, CA 90095-1522
(310) 206-1532

The work reported herein was supported by the National Science Foundation under the Interagency Educational Research Initiative (Award Number 0129406).

IMPROVING FORMATIVE ASSESSMENT PRACTICE WITH EDUCATIONAL INFORMATION TECHNOLOGY

Terry P. Vendlinski, David Niemi, Jia Wang, and Sara Monempour
CRESST/University of California, Los Angeles

Abstract

This report describes a web-based assessment design tool, the Assessment Design and Delivery System (ADDS), that provides teachers both a structure and the resources required to develop and use quality assessments. The tool is applicable across subject domains. The heart of the ADDS is an assessment design workspace that allows teachers to decide the attributes of an assessment, as well as the context and type of responses the students will generate, as part of their assessment design process. Although the tool is very flexible and allows the above steps to be done in any order (or skipped entirely), our goal was to streamline and scaffold the process for teachers by organizing all the materials for them in one place and to provide resources they could use or reuse to create assessments for their students. The tool allows teachers to deliver the assessments to their students either online or on paper. Initial results from our first teacher study suggest that teachers who used the tool developed assessments that were more cognitively demanding of students and addressed the "big ideas" rather than disassociated facts of a domain.

Background

Clearly, most teachers want their students to master the content they teach. Moreover, as both incentives and disincentives in American K–12 education are increasingly tied to student performance (No Child Left Behind [NCLB], 2002), teachers have an even greater impetus to improve student achievement. Improved student learning depends, in large part, on the capabilities of classroom teachers and their ability to encourage conceptual change in student thinking rather than merely attempting to add more factual knowledge to what students already know (Mayer, 2003).

The required capabilities include not only a teacher's own content knowledge, but also the teacher's knowledge of how to teach that content effectively to others (Nathan, Koedinger, & Martha, 2001). Cognitive science has clearly demonstrated that students are not *tabla rasa*; rather, we know that students bring knowledge and mental models that teachers must recognize when designing and implementing instruction (Bransford & Schwartz, 1999; Zull, 2002). Consequently, timely and informative feedback, derived from good formative assessment, would seem to be a critical link in integrating these various strands and improving student learning.

Black and others have demonstrated that formative assessment can dramatically improve student achievement, but only if such assessment guides changes in day-to-day classroom practice (Mayer, 2003; Black & Wiliam, 1998; Wilson, 2004). Such improvements, however, are evident only if teachers and students can understand the information provided by the assessments (Mayer, 2003). Changing teachers' assessment and instructional practices, however, can be difficult. Teachers' preparation in assessment is often non-existent and teachers' content knowledge may be insufficient for a deep understanding of the concepts and principles that they are trying to teach and assess. Nevertheless, the literature (e.g. Frederiksen & White, 2004), and our own experience suggests that one way we can begin to rectify these shortcomings is to let teachers themselves become students of assessment practice. The question this report addresses is: how do we improve the assessment practices of teachers? Anderson (1993) demonstrated that skills are best learned for transfer to a variety of situations when a learner represents the skills as "general rules" rather than fixed responses. Based on more recent cognitive research, Clark and Mayer (2003) suggest an even more focused instructional paradigm:

1. Highlight important information;
2. Minimize the burden on working memory so rehearsal can take place;
3. Integrate new and old knowledge by requiring active processing (e.g. practice exercises);
4. Situate practice of the newly acquired knowledge in a context where it will be used; and
5. Help learners acquire the metacognitive skills necessary for successful learning.

We have integrated the cognitive stages advocated by Anderson (1993), and Mayer and Clark (2003) in four ways. First, we focus the assessment designer's attention on critical aspects of the assessment design process. The essential aspects of assessment design are described by the CRESST assessment model, the organizing principles of domain knowledge, and research in common misconceptions held by students. Second, we use online tools to scaffold the assessment development process and suggest a way to proceed through the design process that logically follows the questions that one must ask to produce quality assessments. Third, a design wizard, tutorial and help are available to the user as they repeatedly develop assessments that they can actually use in their classrooms. Because these assessments and their performance characteristics can be archived in the system, teachers can make incremental changes between administrations of the items, thereby improving item quality by repeatedly building on previous knowledge. Finally, teachers can use the tool to create assessments without any additional help or support.

The Formative Assessment Model

CRESST researchers have conducted extensive experimental research in model-based, cognitively sensitive assessments (e.g., Baker, 1997; Baker & Mayer, 1999; Niemi, 1996) and have moved their research-tested models into large-scale trials in the Hawaii State assessment, the Los Angeles Unified School District assessment program, and in the Chicago Public Schools. In nearly all of our assessment work to date, we have used a model-based approach defined by Baker (1997). Model-based assessment design is an approach to the development of assessments based on the cognitive demands of the task nested within a particular content area (e.g., Klein, O'Neil, & Baker, 1996), and the application of domain-independent specifications that serve as templates for the creation of assessments comparable across different topic or content areas. Baker, Aschbacher, Niemi, and Sato (1992) laid out the specifications for the general approach, which has also been used to develop other assessments in science and mathematics. These assessments provide both formative and summative information in line with the latest thinking in the learning and cognitive sciences (Pellegrino, Chudowsky, & Glaser, 2001).

Another relevant body of research demonstrates the importance of a teacher's conceptual understanding and domain expertise. Understanding of core principles and concepts (the "big ideas") in a subject domain results in more flexible and generalizable knowledge use, improves problem solving, and makes it easier to make sense of and master new facts and procedures (e.g., Gelman & Lee, 1995). If conceptual understanding is essential for high student performance, it is even more critical to teaching for high achievement in science. Assessment design efforts must, therefore, assist teachers in focusing in on the "big ideas", and fine-grained analysis of the types of knowledge and skills that underlie high student performances in science.

A final knowledge base we draw heavily on is the *Facets in Thinking* perspective of Minstrell and others (e.g., Minstrell, 2000). Now in use in several content areas, such as introductory physics, university statistics, middle school mathematics, and environmental science, the facets of students' thinking are individual pieces, or constructions of a few pieces, of knowledge or strategies of reasoning that have been derived from research on students' misconceptions and from classroom observations by teachers.

The Assessment Design and Delivery System (ADDS)

Given that good formative assessment is critical in student learning and that teachers have little training in developing such assessments, our challenge is to provide that

instruction to help teachers (both pre-service and in-service) become skilled at formative assessment practice.

The Assessment Design and Delivery System (ADDS) is a powerful set of tools that (a) provides utilities for individual teachers or teams of teachers to become designers and users of assessments that yield usable information to guide their practice and student learning, and (b) embeds content, assessment, and pedagogical knowledge to assist teachers in both designing assessments and interpreting student progress. ADDS is composed of four tools: the Designer, the Assembler, the Scheduler, and the Gradebook.

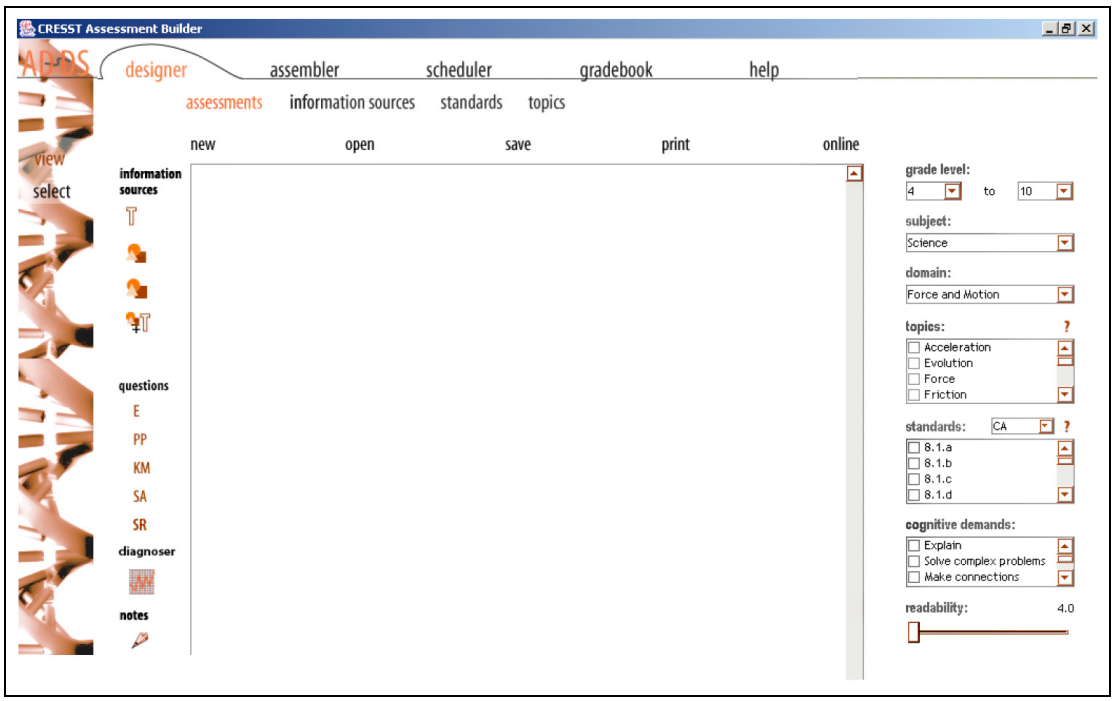

Figure 1. The layout of the Designer in the ADDS.

The **Designer** (see Figure 1) is essential to both assessment development and teacher learning. The Designer scaffolds a teacher-user's thinking about the assessment that will be most useful in a particular situation. Scaffolding serves to both focus a user on the essential attributes of a high-quality assessment and as an aid in searching for exiting assessments. Some of the assessment attributes are commonplace. For example, it is essential that the grade-level and linguistic complexity of the item match the general ability level of the target population. It is, however, the consideration of more atypical attributes of an assessment that the research cited above and our experience suggest are key to developing a teacher-user's assessment acumen. For example, one of the most critical attributes in developing good formative assessment is the need to specify the depth and type of knowledge a student will need to complete a task successfully. For example, the cognitive difficulty of recalling

previously presented data differs greatly from the cognitive difficulty of explaining an idea or constructing an argument. While the ADDS accommodates both types of cognitive demands, it pushes users to distinguish, for example, between deep understanding and mere recognition, and to design assessments appropriately. Another key requirement is specification of the standard or topic to be assessed. While some (e.g. Stiggins, 1994), have argued the need for assessment designers to explicitly state the standard or topic to be assessed for some time, such a requirement is only becoming ubiquitous since NCLB (2002).

The ADDS also allows teacher-users to create a context in which students can actually apply the information. Information presented in the context of problem solving is more likely to be spontaneously used than information presented in the form of simple facts (Bransford & Schwartz, 1999). In ADDS, this is accomplished by introducing complex information sources, which students must interpret by using whatever prior knowledge they have, into the assessment. Information sources can be textual, images, or animation/video/audio files. The ADDS contains a number of these sources, but a teacher can also import any of these types of information sources into our database for private use. The final two aspects an assessment designer must consider when using the ADDS Designer are the question prompt and scoring rubric. We have found that development of a rubric when designing the assessment helps to refine the question or prompt, and possibly the information source. Furthermore, a clear and concise description of expected student responses can improve the quality of other assessment components.

In the **Assembler** a teacher-user can weave together one or more assessments (ones that they have created, those from the data base, or a combination of both) into a "test". We have designed the Assembler to aggregate and show all the topics and standards assessed as a user adds each assessment to a test. Consequently, the user can easily see the breadth of coverage of the entire test as it is built.

The **Gradebook** serves as a record of student achievement as well as functioning as an interface that allows teachers to add student and class information to the ADDS database. When students take an electronically scorable assessment, their scores are automatically entered in the Gradebook, along with each student's total score, class averages of different tests, etc. Pages of the Gradebook can be printed, including students' login names and passwords.

Once assessments have been assembled into tests, a user can schedule tests for online delivery (if desired) by entering the **Scheduler**. Users can specify the date and time of test

delivery as well as the length of time students are allowed to complete the test. Tests may also be printed for pencil and paper administration.

In both usability and actual teacher design studies, we have found that integrated help is essential. The ADDS provides three degrees of user assistance. For users new to both assessment design and the ADDS system, the ADDS includes a tutorial. The **Tutorial** both teaches users the fundamentals of the CRESST assessment design models and explains the ADDS system. For users that have a greater understanding of the basics of formative assessment, the **Wizard** guides the user step-by-step through the process of building or selecting assessments. Compared with the stand alone Designer, the Wizard offers an alternative, more highly structured method of assessment development. Finally, basic **Help Menus** are integrated into ADDS so users can access help by functionality (Contents), by definition (Index), to find answers to frequently asked questions (FAQ), or to search the entire Help contents (Search).

INITIAL RESULTS

The ADDS has successfully completed two usability studies and both teachers and district assessors have designed assessments using the system. In addition, the assessment database contains almost 50 public assessments and a much larger number of individual teacher assessments not yet designated as public. Once individual teacher assessments are actually used in practice, and the data they produced is analyzed, they could also be added to the list of publicly available assessments.

Usability Studies

In December 2003, five science teachers and, at a later time, two CRESST staff were recruited to provide feedback on the interface of the wizard and tutorial portions of ADDS. Teachers were given about 2 hours to go through each interface and were given handouts with screenshots that they could use to make comments on. After each interface teachers were also asked to complete a short questionnaire that asked about features they liked, disliked, or would change about the way the program operates and the ease of use of the graphical interface.

In October 2004, we conducted a second usability study with a group of 16 science teachers. After 2 hours of introduction to ADDS, teachers were asked to create assessments on their own using the Designer in ADDS. Examples of teachers' comments about each interface are presented in the sections that follow.

Designer: The usability study revealed three areas that needed better explanation.

1. Teachers initially found the idea of an assessment task confusing and wanted to know if a task could be multiple questions rather than just a single question. They also wanted to know if there was a way to tie standards to a particular question. We resolved this question by re-labeling "tasks" as "assessments" since teachers seemed more familiar with that term. In addition, we added a more detailed explanation of the term "assessment" in the tutorial, wizard and help. The difficulties experienced by the teachers in the usability study seem to mirror the difficulties found in actual experience. Teachers often want to assess complex ideas with multiple questions, yet research and experience suggests that these questions should each be tied to a particular assessment objective (i.e. a standard or topic). Because ADDS only allows assessments to be so articulated, the teacher is constrained to align assessment questions to address a single topic or standard, or break multiple-question assessments into pieces. These assessments can then be combined into a single test.

2. Teachers were confused by the relationship between cognitive demands and the type of question being posed. This confusion also uncovers important foundational concepts in assessment development. Although some levels of student cognition are easier to assess with certain question types (e.g. explain with an essay task), it is not the case that particular task or question types are intrinsically linked to particular demands; e.g. that multiple-choice items can only assess memorized facts. Our desire is that teachers decide what level of cognitive demand is appropriate to the assessment goal, and to construct or select the assessment accordingly. Toward that end, we augmented Help, Wizard and the Tutorial with more detailed explanations about the relationships between question types and cognitive demands.
3. The final set of questions in the usability study regarded assessment scoring and rubrics. Some teachers did not like the idea of creating a rubric based on expectations of student responses even though they could later refine the rubric based on actual student work over multiple administrations of the test. However, our experience suggests that the process of developing a rubric encouraged teachers to revise assessment questions for clarity and so that they become more clearly focused on the standard or topic being assessed. We also expanded our description of score points and rubrics in the Wizard, Tutorial and Help menus.

Wizard. In general the respondents felt that the wizard interface was straightforward and informative. A couple of the teachers were also interested in measuring competency in other content areas, like math and language arts, to show that the tasks are multidisciplinary and they suggested adding more standards.

Tutorial. The most common suggestion to improve the tutorial was to allow teachers to switch between the tutorial screen and the assessment designer since teachers were required to complete an actual assessment while they used the tutorial. An explanation of how to accomplish this has been added to the current version of the ADDS.

Design Studies

A few general observations can be made from the 33 middle school science teachers that participated in the October 2004 study. At the onset of the study, we asked all 33 teachers to design a test on paper that would adequately assess an individual student's understanding of either genetics (for 7th- and 9th-grade teachers) or motion (for 8th-grade teachers). These concepts appear in the California state standards for the respective grades. We had asked the teachers to bring any materials that would support their test development for the topics they taught during the year with them to the study, and most teachers brought texts and sample tests with them. Teachers were free to use this material as they developed their test questions.

After collecting their assessment questions, we randomly divided the teachers into two groups. Individuals in the control group were given 2 hours to design a second test on paper

that would adequately assess an individual student's understanding of either evolution (for 7th- and 9th-grade teachers) or force (for 8th-grade teachers). Here again, these concepts aligned with the California state standards for the respective grades. Individuals in the treatment groups were given 2 hours of training on the ADDS system and then given 2 hours to design a test in ADDS that would adequately assess an individual student's understanding of either evolution (for 7th- and 9th-grade teachers) or force (for 8th-grade teachers). We collected the assessments from both groups for study. The data from this study supports three broad observations.

"Big Ideas." The teachers in the treatment group were much more likely to begin the assessment development process by noting the broad idea that they were trying to assess, and their assessments were more likely to have the students address these "big ideas" rather than merely recalling specific facts from a particular unit of study. For example, one genetics teacher noted the idea that "If you do not live long enough to reproduce, your genes die off" Without our prompting, teachers were unlikely to develop such focused tests. No teachers in the control group and no teacher prior to treatment apparently used "big ideas" as a basis for test development.

We believe that using the "Big Ideas" as a basis for test development will encourage teachers to develop assessments that allow better inferences about how deeply students understand the important concepts in a field of study.

Rubrics. Only teachers in the treatment group developed rubrics or detailed the responses they expected to receive back from students. While not all of the rubrics were well developed, teachers did consider them as part of their assessment design without being constrained to do so. Here again, we addressed the concept of rubric development only with the treatment group.

Our experience suggests that the very process of rubric development encourages test writers to clarify or refine the test question. Experience also suggests that teachers refine the rubrics as they evaluate student work. Here again, we have witnessed both processes positively impacting assessment practice in schools. Rubrics can also be aids for instructional development and content building for teachers since teachers can now clearly see not only what their students are expected to know, but also how they will be expected to use that knowledge. Unfortunately, most teachers seldom keep their rubrics from one year to the next and so the possibility of long-term assessment "polishing" (National Research Council, 1999; Lewis, Perry, & Hurd, 2004) is lost. Based on research and our own experience, we believe the capability of ADDS to maintain assessment items with their associated, modifiable

rubrics over the course of many years has the potential to significantly improve both instruction and student learning.

Technology. Teachers in the non-treatment group never included video or online sources to prompt student thinking about an assessment question they were designing. In fact, teachers in the non-treatment group seldom used any information sources when designing assessments. One re-plausible explanation might be that teachers designing assessments online were more likely to explore the World Wide Web for information sources or assessment contexts in which students could apply the concepts being assessed. In either case, teachers tended not to use information sources from books or other off-line materials

In general, we found that assessments that included web resources are far more likely to ask students for higher order thinking that those that do not include such resources. Nevertheless, the learning curve among teachers in the treatment group was steeper than expected as evidenced by the amount of time teachers required to construct their assessments. In general, teachers using the ADDS produced fewer assessment questions during the 2-hour design period than when they or their peers designed tests using pencil and paper. Consequently, while the ADDS developed tests appear to be asking students for higher levels of thinking, the cost associated with this improvement appears to be that teachers take time to develop such tests, at least in the period when they are still learning to use the technology

CONCLUSIONS

The research literature and our experience suggests that scaffolding the assessment development process for teachers and providing a means whereby assessments can be continually "polished" should improve the quality of classroom formative assessments. The online ADDS is intended to structure assessment design using a cognitively sensitive, model-based framework designed by CRESST researchers and field tested in large school districts around the United States. In addition, the ADDS incorporates important concepts from the novice-expert literature and research on student misconceptions to enrich the assessment development process. While our initial evaluation of data from design studies noted that, at first, the technology itself increased the time necessary to develop assessments, we also noted that the resulting assessments were often probing student thinking at a deeper level (the "big ideas" of a knowledge domain), included expected student responses and scoring rubrics, and situated the tasks for students in a context where the student could apply knowledge or a concept being assessed. Moreover, experience (both ours and that of others) suggests that as tests are revised based on actual student responses, both the assessment itself and the instruction surrounding the assessment improve. Consequently, we believe that the ADDS has the potential to positively effect assessment practice and student learning in classrooms where it is regularly used by teachers.

REFERENCES

Anderson, J. R. (1993). *Rules of the mind*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Baker, E. L. (1997, Autumn). Model-based performance assessment. *Theory Into Practice, 36*, 247–254.

Baker, E. L., Aschbacher, P. R., Niemi, D., & Sato, E. (1992). *CRESST performance assessment models: Assessing content area explanations* (CRESST Tech. Rep. No. 652). Los Angeles: University of California, National Center for Research on Evaluation Standards, and Student Testing (CRESST).

Baker, E. L., & Mayer, R. E. (1999). Computer-based assessment of problem solving. *Computers in Human Behavior, 15*, 269–282.

Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*(2), 139–148.

Bransford, J. D., & Schwartz, D. L. (Eds.). (1999). *Rethinking transfer: A simple proposal with multiple implications* (Vol. 24). Washington DC: American Educational Research Association.

Clark, R. C., & Mayer, R. E. (2003). *e-Learning and the science of instruction*. San Francisco, CA: Jossey-Bass/Pfeiffer.

Frederiksen, J. R., & White, B. (2004). Designing assessments for instruction and accountability: An application of validity theory to assessing scientific inquiry. In M. Wilson (Ed.), *Towards coherence between classroom assessment and accountability* (pp. 74–104). Chicago: National Society for the Study of Education.

Gelman, R., & Lee, M. G. (1995). Trends and developments in educational psychology in the United States. *Educational studies and documents,* (new ser., no. 61, pp. 23–52). Paris: UNESCO.

Klein, D. C. D., O'Neil, H. F., Jr., & Baker, E. L. (1998). *A cognitive demands analysis of innovative technologies* (Report to ISX/DODEA, CRESST Tech. Rep. No. 454). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Lewis, C., Perry, R., & Hurd, J. (2004). A deeper look at lesson study. *Educational Leadership, 61*(5), 6–11.

Mayer, R. E. (2003). *Learning and instruction*. Upper Saddle River, NJ: Merrill Prentice Hall.

Minstrell, J. (2000). Student thinking and related assessment: Creating a facet-based learning environment. In N. S. Raju, J. W. Pellegrino, M. W. Berthenthal, K. J. Mitchell, & L. R. Jones (Eds.), *Grading the nation's report card: Research from the evaluation of NAEP* (Report of the Committee on the Evaluation of National and State Assessments of Educational Progress, Commission on Behavioral and Social Sciences and Education, National Research Council; pp. 44–73). Washington, DC: National Academy Press.

Nathan, M., Koedinger, K., & Martha, W. (2001, April). *The expert blindspot: When content knowledge and pedagogical content knowledge collide.* Paper presented at the 2001 AERA Annual Meeting, Seattle, WA.

National Research Council. (1999). *Global Perspectives for Local Action: Using TIMSS to Improve U.S. Mathematics and Science Education (Professional Development Guide)*. Washington DC: National Academy Press.

Niemi, D. (1996). Assessing conceptual understanding in mathematics: Representation, problem solutions, justifications, and explanations. *Journal of Educational Research, 89*, 351–363.

No Child Left Behind Act of 2001, 20 U.S.C. § 6301 (2002).

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

Stiggins, R. J. (1994). *Student-centered classroom assessment*. New York: Macmillan College Publishing.

Wilson, M. (Ed.). (2004). *Towards coherence between classroom assessment and accountability*. Chicago: National Society for the Study of Education.

Zull, J. (2002). *The art of changing the brain: Enriching the practice of teaching by exploring the biology of learning*. Sterling, VA: Stylus Publishing, LLC.