# Assessing Inquiry Learning

Diane Jass Ketelhut and Chris Dede
Harvard University Graduate School of Education
diane@brown.alumni.edu

## Abstract

In this paper, we provide an overview of the design of an inquiry-based curriculum project, and then offer a comparative analysis of the outcomes of two methods for assessing student understanding of the inquiry process. Our findings indicate that the complex nature of scientific inquiry is better captured using an alternative method of assessment in addition to a more traditional multiple-choice test.

## Problem

Implementing inquiry in the science classroom is a major emphasis in science education (AAAS 1990, 1993; NRC, 1996). For example, the National Science Teachers Association recently issued a position statement recommending the use of science inquiry as a method to help students understand the processes and content of science (National Science Teachers Association, 2004). However, currently, there is a competing push in science for coverage of material found on state and national standardized tests; in many situations, this competing push forces the emphasis in science classrooms to change from inquiry-based instruction to test-preparation (Falk & Drayton, 2004). Could this dilemma of teaching scientific process versus covering test content be resolved via the inclusion of more inquiry-based questions on these standardized tests? While this may provide teachers and schools with incentives to cover inquiry skills as well as factual content, this solution raises a different concern: Can learning from good inquiry-based projects be adequately assessed using a standardized test format? What kind of assessments will allow valid inferences about whether a student has learned how to engage in inquiry, particularly in the "front end" inquiry processes used to derive a strategy for making sense out of complexity: problem finding, hypothesis formation, experimental design?

Using an NSF-funded Multi-User Virtual Environment (MUVE) as a pedagogical vehicle, our research team is exploring how a technology-intensive learning experience that immerses participants in a virtual "world" whose citizens face chronic illnesses can help middle school students learn both deep inquiry skills and science knowledge. In this paper, we provide an overview of the design of this inquiry-based curriculum project. We then offer a comparative analysis of the outcomes of two methods of assessing student understanding of the inquiry process in order to clarify the extent to which typical forms of test items can validly measure students' inquiry skills.

## Theoretical Underpinnings

### *Inquiry*

What is "inquiry?" The range of possible responses to this question is large. Some refer to inquiry as a set of process skills that include questioning, hypothesizing and testing while others equate it to "hands-on" learning. The National Science Education Standards (NSES) define scientific inquiry as "the diverse ways in which scientists study the natural world and propose explanations based on the evidence derived from their work…also …the activities through which students develop knowledge and understanding of scientific ideas, as well as an

understanding of how scientists study the natural world" (National Research Council, p 23). The standards go on to explain that scientific inquiry is:

> a multifaceted activity that involves making observations; posing questions; examining books and other sources of information to see what is already known; planning investigations; reviewing what is already known in light of experimental evidence; using tools to gather, analyze, and interpret data; proposing answers, explanations, and predictions; and communicating the results. (National Research Council, 1996, p 23).

*River City, a MUVE*

Our project studies how a technology-intensive learning experience that implements problem-based inquiry science curricula can provide both deep inquiry skills and content coverage. In particular, we are working to dramatically improve the educational outcomes of the bottom third of students, pupils who even by middle school often have given up on themselves as learners. These students are disengaged from schooling and typically are difficult to motivate even by good teachers using conventional inquiry-based pedagogy. We are investigating whether educational Multi-User Virtual Environments (MUVEs), which resemble the entertainment and communication media that students use outside of school, can reengage them in learning. MUVEs enable multiple simultaneous participants to access virtual contexts, to interact with digital artifacts, to represent themselves through "avatars," to communicate with other participants and with computer-based agents, and to enact collaborative learning activities of various types. This last we use to create a community of inquiry learners.

Our "River City" MUVE is centered on the NSES inquiry skills listed above, as well as on content related to national standards and assessments in biology and ecology. The virtual "world" consists of a 19$^{th}$ century city with a river running through it, different forms of terrain that influence water runoff, and various neighborhoods, industries, and institutions such as a hospital and a university. The students themselves populate the city, along with computer-based agents, digital objects that can include audio or video clips, and the avatars of instructors. Content in the right-hand interface-window shifts based on what the participant encounters or activates in the virtual environment.

In River City, students work in teams to develop hypotheses regarding one of three strands of illness in the town (water-borne, air-borne, and insect-borne). These three disease strands are integrated with historical, social and geographical content, allowing students to experience the inquiry skills involved in disentangling multi-causal problems embedded within a complex environment. Once their analysis is completed, students write an authentic lab report in the form of a letter to the Mayor of River City, delineating their hypothesis, experimental design, findings and recommendations for solving the city's health problems. At the end of the project, students compare their research with other teams of students in their class to outline some of the many potential hypotheses and causal relationships embedded in the virtual environment. Complete details of the project have been reported previously (Nelson, Ketelhut, Clarke, Bowman and Dede, 2005).

*Inquiry and River City*

In River City, students engage in all aspects of inquiry as defined by the NSES. These aspects are listed below, and we have mapped each onto where in River City the behavior can be observed:

1. "Making observations" – students move around the world, making visual and auditory observations about the city and its inhabitants.
2. "Posing questions" – students can ask a question of the computerized residents of River City and elicit information that often offers a clue about the problems.
3. "Examining books and other sources of information to see what is already known" – students can access information from books in the River City library as well as from guidance hints, embedded clues in digitized historical images, and the hospital admissions record.
4. "Using tools to gather, analyze, and interpret data" – students can gather data from two tools: a water sampling tool and a 'bug-catching' tool (see figure 2). Each tool is activated by a student click to draw a sample; the student then counts bacteria in a microscope-like screen.
5. "Planning investigations"—students are guided through a generalized process of the scientific method, culminating in creating a unique experiment to test their hypothesis about the problems in River City. They not only design the hypothesis, but also the procedure and data-collection methodology.
6. "Reviewing what is already known in light of experimental evidence"—students spend over a week in River City gathering evidence on the problem from multiple sources, including embedded experts in the form of hospital doctors and university researchers, prior to conducting their own experiments. They must use that information to design their experiments. Once they have analyzed the results of their own experiments, they must compare it with what they hypothesized earlier and to what the embedded experts told them. This process is made transparent in the performance assessments.
7. "Proposing answers, explanations, and predictions"—students create a hypothesis based on collecting evidence to predict what they think is causing a piece of the problem in River City. They re-evaluate that hypothesis in the light of the results of their experiment.
8. "Communicating the results"—at the end of the project, students take part in a classroom-based research conference, delineating their thinking, experiment and results. During this conference, students try to piece together all student results to understand the larger picture of what is making everyone ill in River City.

Students work in teams to gather data, develop hypotheses regarding one of three strands of illness in the town (water-borne, air-borne, and insect-borne) and then to test their hypothesis. These three disease strands are integrated with historical, social and geographical content, allowing students to experience the inquiry skills involved in disentangling multi-causal problems embedded within a complex environment. After testing their hypothesis, students analyze their data using graphs and tables and then write an authentic lab report on their findings in the form of a "Letter to the Mayor of River City." Finally, at the end of the project, students compare their research with other teams of students in their class to delineate the many potential hypotheses and causal relationships embedded in the virtual environment.

*Assessment of inquiry*

Since inquiry involves higher order thinking skills that are not easily measured with multiple-choice tests (Resnick & Resnick, 1992), we chose to design a traditional (pre/post

surveys) and an alternative (our Letter to the Mayor) method of assessment of student learning in our MUVE. Published reviews are available that detail the merits of each of these styles of assessments (Mehrens, 1998; Kane, Khattri, Reeve & Adamson, 1997; Moore, 2003). To summarize: Proponents of alternative assessments view them as capturing student understanding better than standardized tests, which they feel measure decontextualized knowledge. Opponents argue that performance-based assessments are not cost effective, cannot be compared from teacher to teacher due to individual grading differences, and are inconclusive about what the tasks are actually measuring (Stecher & Klein, 1997). This paper will examine this debate from the perspective of our project.

## Design and Procedure

*Research questions*

1. Is there a difference in inquiry learning between the different treatments, as evidenced by the Letters to the Mayor?
2. Do the pre/post surveys and Letters to the Mayor show similar patterns of inquiry learning?

*Student Population*

This papers present partial results of two sets of implementations conducted in the Spring and Fall of 2004 in urban and rural areas in New England, the Southeast and the Midwest. The student population of 1660 students in these areas had high proportions of ESL and free-and-reduced lunch pupils. A total of 8 schools, 12 teachers and 61 classes are involved in this analysis.

*Procedures*

In order to study the effect of learning theory on student outcomes, we developed three variations of the River City curriculum for these implementations. Variant GSC centers on a guided social constructivist (GSC) model of learning-by-doing, in which inquiry experiences in the MUVE, supported by both virtual and physical lab notebooks, alternate with in-class interpretive sessions led by the teacher. Another variant shifts the learning experience to a situated pedagogy based on expert modeling and coaching (EMC), in which students interact with expert avatars and agents embedded in the MUVE. The third variant, Legitimate Peripheral Participation (LPP), is based on Lave and Wenger's (1991) concept of a community of practice; students move from simple peripheral roles to more complex tasks through tacit forms of learning such as internships. These three River City variants were compared to a "control" condition that utilized a paper-based curriculum in which the same content and skills were taught in equivalent time to comparable students without using computers, via a guided social constructivist-based pedagogy. The control curriculum (EI) included features similar to River City, such as a historical scenario and unknown disease transmission. In addition to experimental design and analysis, this curriculum also included physical experimentation. This type of control curriculum enables us to focus on the strengths and limits of MUVEs, as well as the types of pedagogy best supported by this medium.

During the Spring 2004 implementation, students were assigned randomly within class to one of the two River City treatments, GSC and EMC, with teachers instructed to minimize cross-contamination of treatments. In the Fall 2004 implementations, students were randomly assigned

within class to one of three River City treatments, GSC, EMC and LPP. In all implementations, each teacher also taught the control curriculum, EI, which was randomly assigned to one class, except in one treatment site where the teacher taught three classes of EI to students who had gone through the River City curriculum the previous year (those students are not included in this analysis).

River City incorporates an underlying database that captures individual student activity in the MUVE with a timestamp, allowing us to analyze students' microbehaviors (such as where students went, who they talked to, what they said) throughout the implementation. After designing and conducting their experiments, students in both the control and River City treatments were asked to write letters to the Mayor in which they discussed their hypothesis, experimental design, results and recommendations for solving the city's health problem.

Both qualitative and quantitative data were collected from students and teachers over the three-week implementation period. Pre- and post-intervention, the students completed an affective pre-survey. To assess understanding and content knowledge (science inquiry skills, science process skills, biology), we administered a content test, pre- and post-intervention, of our own design.

To support teachers, we conducted an extensive professional development program, delivered both face-to-face and online. This program focused on content review, alternative pedagogical strategies based on different theories of learning, facilitation strategies while students are using the MUVE, and interpretive strategies for leading class discussions. The teachers collected demographic data and rated their expectations of students' successes and motivation with the project. Teachers responded to a pre- and post-questionnaire regarding their methods, comfort with technology, and reflections on using the MUVE in their science class.

*Measures*

Affective pre and post survey:

This measure was adapted from three different surveys, Self-Efficacy in Technology and Science (Ketelhut, 2005), Patterns for Adaptive Learning Survey (Midgley, et al, 2000), and the Test of Science Related Attitudes (Fraser, 1981). This modified version has scales to evaluate students' science efficacy, thoughtfulness of inquiry, science enjoyment, and career interest in science. Its individual subscales have reasonable internal consistency reliability estimates (ranging from .8 to .93), as well as validity evidence from prior research (Ketelhut, 2005).

Content pre and post surveys:

This measure was our own design. Questions assessed student knowledge in two main categories: scientific inquiry and disease transmission. The inquiry portion has a range of 0→17. Internal consistency reliability was estimated using Cronbach's alpha and Principal Components Analysis, indicating a reliability of .86. Content validity was established through analysis by a team of experts.

Alternative assessment:

As indicated earlier, at the end of the project, students wrote a letter to the Mayor of the town as a performance of understanding assessment. These were coded by a rubric developed through multiple iterations by the team with possible scientific inquiry subset scores ranging from 0→26. Student letters were coded for:

- Front end inquiry skills of identifying a problem and forming a hypothesis (*frontend),*

- Evidence of analysis of collected data, including tables and graphs and results (*analysis),*
- Discussion of collected data including referring back to the original hypothesis (*explanation*),
- Listing future areas of research (*prediction*),
- Understanding the multivariate nature of the problem (*multivariate*),
- Experimental design (*designing experiment),*
- Coherence between the chosen problem, hypothesis, experimental design and analysis (*coherence),* and
- Overall understanding of all pieces of scientific inquiry as outlined above (*scientific inquiry overall*).

## Findings

The quantitative data were analyzed with SAS. Descriptive statistics, correlations and multi-level modeling regression with class as the cluster variable models were run. Checks for linearity, normality and homoscedasticity were performed at intervals. No clear violations were noted.

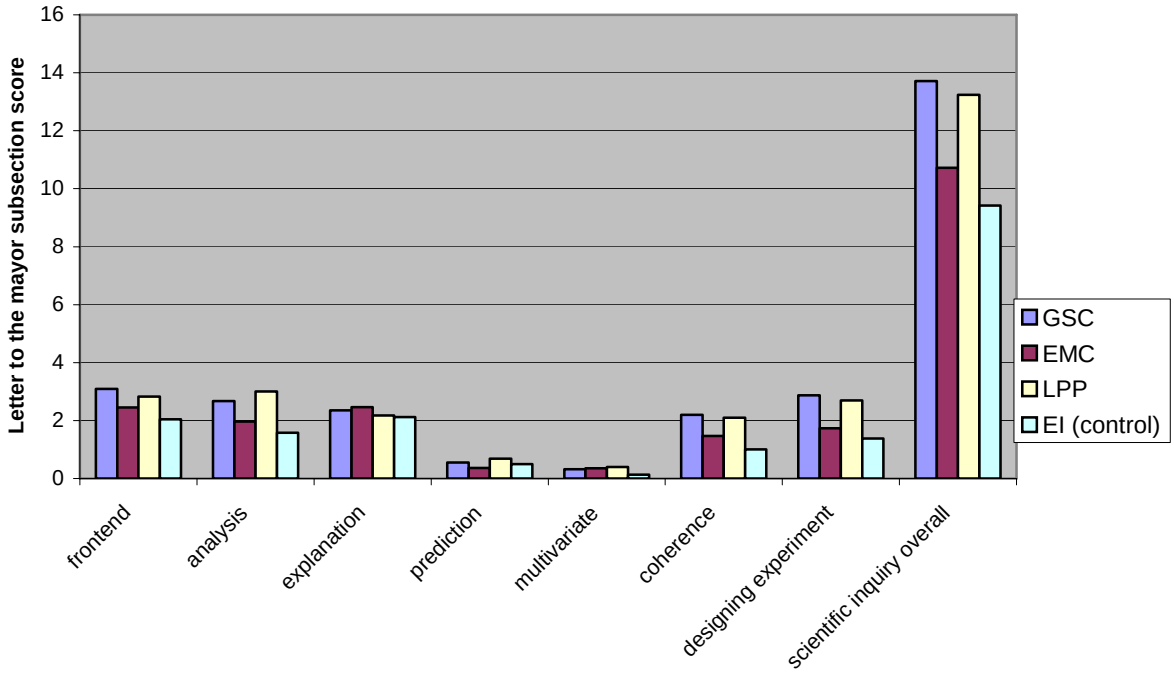
*Results*

Research Question 1:

Our first research question asks whether the Letters to the Mayor show evidence of inquiry that differ by treatment. In our first implementation in Spring 2004, instructions varied somewhat between the River City curriculum and the EI (control) curriculum; as a result, detailed comparison of the letters between treatments for this implementation may not be productive. Therefore, we looked for similar demonstrations of student understanding of the processes of inquiry and for motivation. The letters written for the control curriculum often: were much shorter in length, did not demonstrate motivation or engagement, did not mention the experiment, and did not explicitly recognize the interconnectedness of the chosen problem with other possible causes of the larger problem. Analysis of the letters' evidence of inquiry found that students taking part in the MUVE-based curriculum earned scores more than double ($p<.01$) that of their paper-based control peers, on average (Ketelhut, Clarke, Dede, Nelson, & Bowman, 2005).

For the next implementations in the Fall 2004, the instructions were identical between all treatments, which allowed more detailed comparison between the letters. Indeed, in a random subsample (n=202) of letters coded to date, we did find differences of inquiry learning by treatment, with the paper-based control treatment showing the lowest evidence of inquiry learning. Figure 1 shows average values for the inquiry subscales for each treatment.

**Figure 1. Average scores on Letters to the Mayor, controlling for treatment (n=202).**




Based on average values alone, we can see in Figure 1 that students in the control treatment perform more poorly than students in the River City treatments across the subsections.

To confirm this brief overview and to determine whether these differences were statistically significant, we conducted a multi-level regression analysis of the data, controlling for class-level variations. The results for this analysis confirm the overview seen in Figure 1. EI students did not show the same level of detailed evidence of scientific inquiry in their letters as did students in River City treatments, on average.

Specifically, students in the guided social constructivist treatment (GSC) wrote letters that were more likely than EI students to:

- Show evidence of front end inquiry skills ($p<.06$),
- State a testable hypothesis ($p<.02$),
- Understand the multivariate nature of the problem ($p<.08$),
- Match the pieces of the experimental design to each other ($p<.07$),
- Describe their experimental design ($p<.05$),
- Include tables ($p<.01$).

Similarly, student in the legitimate peripheral participation (LPP) treatment wrote letters that were more likely than EI students to:

- Show evidence of front end inquiry skills ($p<.09$),
- State a testable hypothesis ($p<.07$),

- Understand the multivariate nature of the problem ($p<.01$),
- Match the pieces of the experimental design to each other ($p<.08$),
- Describe their experimental design ($p<.05$),
- Include tables ($p<.05$).
- Show excitement about the inquiry process ($p<.06$).

Finally, EMC students were more likely than EI students to:

- Describe the evidence that they used to formulate their problem ($p<.04$),
- Understand the multivariate nature of the problem ($p<.04$).

Table 1 summarizes these results. Column 1 of this table lists the letter-evaluation categories that showed differences across the treatments. The treatments with significantly higher scores for that category are denoted with a '✳' in columns 2-5; treatments which had worse scores are denoted with a '—' in those same columns.

Table 1. Coded areas of the "letters to the mayor" that showed significant differences ($p<.10$) by treatment in student scores relative to one or more of the other treatments (n=202).

| **Areas that differed significantly by treatment (p<.10)** | **GSC** | **EMC** | **LPP** | **Control** |
|---|---|---|---|---|
| Front end inquiry skills | ✳ | | ✳ | — |
| State a testable hypothesis | ✳ | | ✳ | — |
| Describe experimental design | ✳ | | ✳ | — |
| Show evidence for their problem | | ✳ | | — |
| Show coherence in the pieces of the scientific inquiry process | ✳ | | ✳ | — |
| Summarize data in tables | ✳ | | ✳ | — |
| Understand the multivariate nature of the problem | ✳ | ✳ | ✳ | — |
| Use evidence to justify problem choice | | ✳ | | — |

*Key:* ✳ = Treatment that on average had significantly higher scores relative to the treatment with "-"
— = Treatments that on average had worse scores in this category relative to ✳treatments
= Treatments that on average were not significantly different from the others in this category

As can be seen in Table 1, students in the GSC and LPP treatments had higher scores in nearly every category, whereas students in the control treatment did not do significantly better on any aspect of the letters to the mayor than did the River City treatment students.

Research Question 2:

Research question 2 asks whether our pre/post multiple-choice measure shows the same pattern of understanding inquiry across treatments as our performance assessment. In our first implementation in Spring 2004, improvements were seen across the board on the surveys for knowledge and application of scientific processes; control students improved slightly more than the other two groups: 20% for the control, 18% for the GSC group and 16% for the EMC group (Nelson et al, 2005). This pattern of difference between the treatments varies from what was seen in the analysis of the letters, above.

This was also the case in the Fall implementations. Results from the post survey indicate only one area that differs between the treatments: that of creating a testable hypothesis. As was the case for the Letter to the Mayor, students in the GSC and LPP treatments were more able to create a testable hypothesis, on average, than EI students ($p<.01$). While this confirms one aspect of the analysis of the Letter to the Mayor, it is the only area where there is agreement between the two forms of assessment.

*Comparison of the two methods of assessment*

The above analysis suggests that science inquiry post-test measures blur the distinctions between treatments and do not capture students' understanding of inquiry as well as the Letters to the Mayor. For example, students who scored low on the science inquiry post-test wrote letters that were of similar quality to those written by students who scored higher on the post-test. In addition, in their letters both low- and high-performing students demonstrated a clear causal relationship between the problem and the reason(s) for the problem. As another illustration, in their letters low-performing content students matched the high-performing content students around criterion of stating an opinion regarding the cause of the problem and/or the outcome of the experiment. Interestingly, more of the lower-performing test students met the criteria of providing suggested interventions or further research than students who scored higher on the inquiry test questions. This suggests that the complexity of the MUVE treatment creates intricate patterns of learning more appropriately measured with an authentic activity, such as writing an experimental report. If these results can be generalized to other inquiry activities, this brings to question whether inquiry can be assessed with standardized tests, and if not, what effect this will have on its integration into the standards-based classroom.

## Conclusion

Scientific inquiry is a difficult construct for teachers to implement without support, and the current emphasis on content coverage via high stakes tests often reinforces presentational pedagogies. To support teaching of good inquiry-based curricula, one possible solution would be to increase the emphasis of inquiry-based questions on the current model of standardized tests. Our results indicate, however, that a multiple-choice format is much less sensitive to differences in learning of inquiry than an alternative assessment format. Our recommendation, therefore, would be to encourage the use of multiple forms of assessment to evaluate inquiry.

## References

American Association for the Advancement of Science. (1990). *Science for All Americans*. New York, N.Y. Oxford University Press.

American Association for the Advancement of Science. (1993). *Project 2061: Benchmarks for science literacy*. New York: Oxford University Press.

Falk, J., & Drayton, B. (2004). State Testing and Inquiry-based Science: are they Complementary or Competing Reforms? *Journal of Educational Change, 5*, 345-387.

Fraser, B. (1981). *TOSRA: Test of Science Related Attitudes*. Australian Council for Educational Research, Hawthorne, VIC.

Kane, M., Khattri, N., Reeve, A. & Adamson, R. (1997). *Studies of Education Reform: Assessment of Student Performance*. U.S. Education Department's Office of Educational Research and Improvement, Washington, D.C.

Ketelhut, D. J. (2005, April 4-8). *Assessing Science Self-Efficacy in a Virtual Environment: a Measurement Pilot*. Paper presented at the National Association of Research in Science Teaching Conference, Dallas.

Ketelhut, D. J., Clarke, J., Dede, C., Nelson, B., & Bowman, C. (2005, April 4-8). *Inquiry Teaching for Depth and Coverage via Multi-User Virtual Environments.* Paper presented at the National Association for Research in Science Teaching, Dallas.

Ketelhut, D. J., Dede, C., Clarke, J., Nelson, B., & Bowman, C. (in press). Studying Situated Learning in a Multi-User Virtual Environment. In E. Baker & J. Dickieson & W. Wulfeck & H. O'Neil (Eds.), *Assessment of Problem Solving Using Simulations*: Lawrence Erlbaum Associates.

Lave, J. & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. New York, NY: Cambridge University Press.

Mehrens, W. (1998). Consequences of Assessment: What is the Evidence? *Education Policy Analysis Archives, 6(13).* Online: *http://epaa.asu.edu/epaa/v6n13.html*.

Midgley, C., Maehr, M. L., Hruda, L. Z., Anderman, E., Anderman, L., Freeman, K. E., Gheen, M., Kaplan, A., Kumar, R., Middleton, M. J., Nelson, J., Roeser, R. & Urdan, T. (2000*). Manual for the Patterns of Adaptive Learning Scales (PALS),* Ann Arbor, MI: University of Michigan.

Moore, Wayne. *(2003).* Facts and Assumptions of Assessment: Technology, The Missing Link. *T H E Journal 30 (6).* 20-26.

National Research Council (1996). *National science education standards*. Washington, DC: National Academy Press.

National Science Teachers Association. (2004). *NSTA Position Statement: Scientific Inquiry* (Draft), [Internet]. NSTA. Available: http://www.nsta.org/main/forum/showthread.php?t=1175 [2004, August 9].

Nelson, B., Ketelhut, D. J., Clarke, J., Bowman, C., & Dede, C. (2005). Design-Based Research Strategies for Developing a Scientific Inquiry Curriculum in a Multi-User Virtual Environment. *Educational Technology*, 45(1), 21-27.

Resnick, L.B. & Resnick, D.P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. Gifford & M. O'Connor (Eds.), *Changing Assessments: Alternative Views of Aptitude, Achievement, and Instruction.* Norwell, MA: Kluwer Academic Publishers, 37-75.

Stecher, B. M. & Klein, S. P. (1997). The cost of science performance assessments in large-scale testing programs. Educational Evaluation and Policy Analysis, 19(1), 1-14.