Considerations for the Development of a Preparation for Future Learning Assessment

Drue Gawel, Rachel Phillips, University of Washington, 312 Miller, Box 353600, Seattle, WA, 98195
Email: djgawel@u.washington.edu, rachelsp@u.washington.edu
Vanessa Svihla, University of Texas-Austin, 518 Sanchez, Austin, TX, 78705, vsvihla@hotmail.com
Nancy Vye, John Bransford, University of Washington, 312 Miller, Box 353600, Seattle, WA, 98195
Email: nancyvye@u.washington.edu, bransj@u.washington.edu

Abstract: This paper presents initial findings from on-going research on the development of assessments that measure how students are prepared for future learning in fast changing environments. Issues of assessment are extremely important for measuring the quality of learning and teaching. Typical assessments tend to tell us what students have learned in the past but not necessarily how prepared they are to learn in the future. With this problem in mind, we have begun research on how students are able to use resources (technology-based access to relevant information, social networks, simulations) in order to learn to solve problems. Our goal is to provide more valid measures of students' existing strengths as well as skills and knowledge that they need to learn. Our assessments are designed to be both formative and summative.

Introduction

Thanks to a partnership with Sam Houston and the Partnership for 21St Century Skills in the state of North Carolina, members of the LIFE Center (Learning in Informal and Formal Environments) have had the opportunity to collaborate with North Carolina educators to improve assessments of twenty first century learning. We are beginning in the area of high school science, and striving to align our work with new sets of 21st Century Skills (Partnership for 21st Century Skills, 2003). These include opportunities for students to learn collaboration, multimedia communication, research, creativity, and design. Additionally, the 21st Century Skills require a foundation of global awareness, as well as literacy in the following areas: finance, business, economics, civics, and media.

Great strides have been made with respect to teaching practices, but change is still needed in the structure, function, and resulting value of assessment practices. Many state assessments are more likely to emphasize explicit and factual knowledge than problem solving and thinking skills. Educational leaders in North Carolina worried that their current assessment systems were inadequate for the task of assessing 21st Century Skills.

Framing

North Carolina's goal is to transform their current accountability system into one that assesses and helps students learn 21st Century Skills. Such an accountability system will encourage the teaching of these skills in classroom practices. Our task was to develop a multimedia assessment prototype that would provide formative and summative feedback and include opportunities for students to learn as they took their assessments. A major goal was to help students learn while they were being assessed so that no instructional time was lost to testing. (Sam Houston, Guiding Vision handout, October 23, 2006). The charge and funding for our research originated with Sam Houston, President of the North Carolina Science, Mathematics and Technology (SMT) Center (Burroughs-Wellcome Fund). Dr. Houston connected us with the policy makers and educators at all levels of government within North Carolina, where there is appreciable support.

Rationale for this type of Assessment

Before presenting the details of this particular tool and pilot study, we would first like to explore some of the questions, background research, and learning theory that guided the creation of this tool. Broadly, we wondered, "How can Preparation for Future Learning (PFL) multimedia instructional technology assess 21st Century Skills?" More specifically, we asked, "Do students demonstrate valuable skills in an interactive PFL task that non-interactive multiple choice tests don't reveal?" In a PFL task, students demonstrate their capacity to solve problems by having the opportunity to learn while performing the task in knowledge rich environments (Bransford & Schwartz, 1999). We wanted to design an assessment that could be used in a formative manner (Black & Wiliam, 1998). However, we also want to develop new tools for summative assessments. And our goal is to assess skills in the context of dealing with important content (see Figure 1).

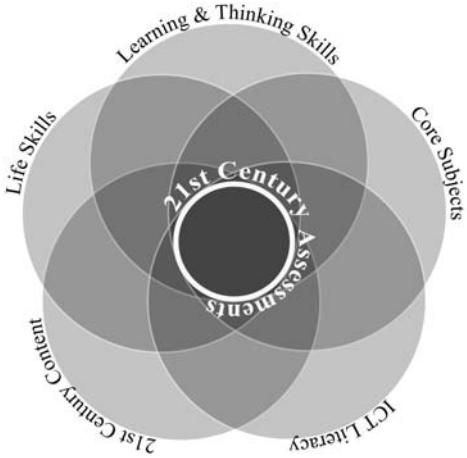

Figure 1. 21st Century assessment lies at the nexus of 21st Century Skills.

In contrast to the need to incorporate and emphasize 21st Century Skills in education, current testing has sometimes moved educational systems in the opposite direction. In a recent study, Au (2007) finds that high-stakes testing increases the direct instruction of fragmented factual knowledge. Bloom's taxonomy has been misinterpreted (in our opinion) and used to justify such practices based on the idea that higher order thinking depends on the prior memorization of facts. In a retrospective of Bloom's Taxonomy, Anderson, Sosniak and Bloom (1994) described this unintended consequence of the Taxonomy and note: "past research has demonstrated that as higher mental processes are emphasized and taught, lower level skills can be learned concomitantly (1994, p. 8)."

Some assessments emphasize problem solving and critical thinking skills. For instance, the Programme for International Student Assessment (PISA) emphasizes extrapolation and application of what students have learned in new situations as opposed to reproducing knowledge. Findings have shown that this particular test more accurately predicts students' future success (Schleicher, 2007). Past work on authentic assessment has had varied implementation success. And although political and technical issues have hindered their success, some have failed due to the high price of implementing these types of assessments on a large scale (M. Wilson & Sloane, 2000). The California Learning Assessment System (CLAS) provided a rich means of evaluating the effectiveness of schools through matrix sampling. Unfortunately, the political climate changed such that every student was required to be tested. Prior to the requirement to assess all students, only 42% of the open ended questions were actually scored (S. M. Wilson, 2003).

New technology provides a potential solution to the cost issue. Tools such as Calipers, and Principled Assessment Designs for Inquiry (PADI), seek to accomplish some of the objectives in our assessment. Calipers uses simulations to assess complex science concepts (Quellmalz, 2007). PADI is a development tool used to create evidence based assessments of student inquiry (Mislevy & Riconscente, 2005). PADI is just beginning to be used in projects in a formative way (Means, 2006). We are hopeful that the depth and complexity of an instructional and assessment tool needed to measure 21st Century Skills can be built leveraging technology such as PADI, Calipers, and Latent Semantic Analysis among others.

Previous work on collaboration in project based learning has emphasized the importance of specific principles for designing curricula: defining learning-appropriate goals, providing scaffolds, ensuring multiple opportunities for formative self-assessment, and supporting the development social roles (Barron et al., 1998). Although we have begun our initial work primarily with individuals, our intent is to create a learning and assessment system that can also be used in collaborative ways.

Methods

This project involves collaboration with various governmental and non-governmental organizations. To show that PFL tasks require 21st Century Skills not demonstrated in multiple choice questions, high school students enrolled in biology at a North Carolina magnet (Medical Science) high school were provided with both multiple choice questions and a PFL challenge scenario. This within-subjects design allows us to compare the variety of skills used by the same student on different assessment tasks.

All students (N=24) completed the multiple choice questions and worked on the PFL challenge during a 90 minute session (see Table 1). Half of the students took the assessments prior to receiving traditional instruction in genetics and the rest took the session post-instruction (instruction consisted of three to five 90 minute periods). Follow-up interviews were conducted with the pre-instruction students a week later to discuss effects the assessment session had on subsequent learning.

Table 1: Study design and sample sizes.

	Treatment		
Pre-instruction group (n=12)	Assessment session (90 min)	Traditional Instruction	Post-Instruction Interview (40 min)
Post-instruction group (n=12)		Traditional Instruction	Assessment Session (90 min)

Biology was chosen as the content area in which to develop the initial instructional tool, in part on recommendation of educators in North Carolina's Department of Public Instruction. We developed a PFL scenario and related multiple choice questions in the context of genetics. These were reviewed by content area experts. The multiple choice questions were constructed by a test question analysis consultant to a large testing corporation and by the classroom teacher. Multiple choice questions were given before and reviewed after the PFL assessment as a way to contrast the learning captured by each. We hypothesized that the challenge might help students see the relationship between the concepts in the multiple choice questions and the task.

In contrast to the multiple choice tests, the PFL scenario placed the student subjects in the role of genetic counselors. The assessment session (see Figure 2) began with students answering the multiple choice questions and rating their confidence level for each question.

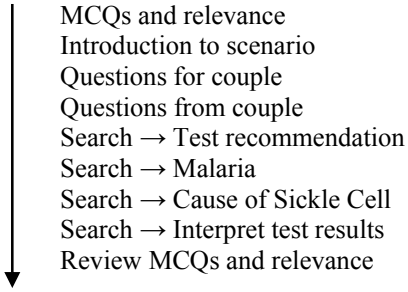

Figure 2. Sequence of PFL task events.

Students were then asked why it might be important to know the information contained in the questions. Students then began the PFL assessment by reading statements from an (imaginary) couple considering having a child, but concerned the child might inherit sickle cell disease. Sickle cell disease was chosen because it is a widely used topic in high-school biology curricula around the country, and is sufficiently nuanced, providing many layers: inheritance, evolution, environmental interaction with genes, political policy and ethics. Additionally this context has been studied elsewhere (Bell, Bareiss, & Beckwith, 1993). Students were asked to take on the role of genetic counselor, which makes the problem more student-centered and eventually, student-driven. Learning theory posits that this is a way to keep students engaged with the learning experience, resulting in more successful learning (Bransford, Brown, & Cocking, 2000). Students were given information (age, ethnicity, family disease history) about the couple in a patient history form, and asked questions by the couple. After reviewing the patient history, students were asked to take on their first task as a genetic counselor: discussing initial thoughts about the scenario. The students were then given time to search on the internet for whatever information they felt they needed to provide the couple with appropriate advice. It is important to point out that the selection of questions and tests for the couple is not a straightforward process that can be found by a simple search on the internet. Then, students developed a list of questions for the couple and recommended appropriate tests that the couple might undergo. Students who satisficed quickly were asked additional questions from the couple and were given time to search so that they could provide the couple with more answers. These follow-up questions redirected focus from Sickle Cell as a collection of symptoms to the interaction of Sickle Cell and Malaria, a focus on natural selection. Following the scenario questions, students were again to review the previous multiple choice questions and their relevance. This concluded the assessment session.

Students who took the assessment prior to instruction, returned the following week for 40 minute interviews. They were asked about how they related to the curriculum taught given their experience with the PFL assessment and if and how it impacted any of their experiences outside of school.

Various data were collected during this pilot study: demographic information; student responses to the multiple choice questions and to the PFL task, both of which the students completed in a research methodology called a "Think aloud protocol," in which students completed while thinking aloud as they answered the questions; the internet search conducted by the students as they worked on solving the problem presented in the task; a before and after query about the usefulness and purpose of the multiple choice questions; post-task

discussion and reflection about assessment and learning with the students; and interview data. The qualitative data were coded by contrasting the scenario and MC questions as well as using grounded theory to allow codes to emerge naturally from the data (Glaser & Strauss, 1967). Responses to the MC questions are contrasted with student reported confidence levels.

Results

Analysis of the study data is ongoing, but we will highlight preliminary findings, which have revealed some consequential outcomes, both for understanding how students engage with traditional assessments and for the development of PFL Assessments. These include student thinking during multiple choice questions, evidence of preconceptions, and more general findings that support the use of PFL assessments, including evidence of student learning.

Student Thinking on Multiple-Choice Questions.

Student thinking on the multiple choice questions was more involved than anticipated. The think aloud protocol captured the deep thinking students engaged in when completing multiple choice questions. The rich data show that students not only have an advanced set of standardized test-taking skills, but also have logical problem solving abilities. This high level problem solving did not always lead to the correct answer on the test, but the data show much more than that of the typical standardized assessment. This implies that current assessments are giving us an impoverished view of student's real capabilities, supporting the assertion that PFL assessments give students the opportunity to show their strengths and abilities, as think aloud protocols are not suitable for large-scale assessment.

One of the problems with MC tests is that students may select the correct answer, yet not be able to apply the concept. Such false positives are evidenced in our research: Although this student got multiple choice questions associated with this concept correct, when explaining how a trait confers Sickle Cell Disease, she is hesitant and questions her factual knowledge: "It - genes - how to word this? It's in your gene and that affects your trait. The trait makes you have… I don't know. I don't think that really makes sense." Another student was highly confident about her (correct) answer: "I would say A because of mutation…A mutation is just like if you have, like if you have a, like if you have long hair, the person might have long hair to come out too." Not only would her misconception about mutations not be discovered in typical testing situations, her lack of understanding would also be strengthened by her sense of confidence and the validation of answering the question correctly.

For each multiple choice question, students were asked to indicate how confident they were in their response. We analyzed these data for mismatches in confidence. We posit that if a student were highly metacognitive, her/his confidence level would be very closely related to correct/incorrect answers. (We will discuss counter-examples in our findings, but take this as a straw-man ideal). We therefore coded the correct answers as "3" and the incorrect answers as "1" and compared these to levels of confidence (1=low, 2=medium, 3=high). We then calculated the mismatch between scores, assuming that a correct answer ought to correspond to high confidence and an incorrect answer to low confidence if the student were highly aware of his or her own understanding. Note that this analysis provides an overestimate score and an underestimate score for students who exhibit both behaviors across questions. The highest possible mismatch score for each question is 2, and given 13 questions answered by each student, the highest possible total mismatch score for a student is therefore 26. In most cases, this is partitioned into underestimates and overestimates, though there is variance across students (22% to 100% of the total mismatch score partitioned as overestimate, with 13 of 24 students favoring overestimating their knowledge). Given 24 students, the highest mismatch score for a question is 48, again, generally partitioned into overestimates and underestimates of their knowledge. On average, there is no difference in students' estimates before and after instruction. Students on average, tend to overestimate their knowledge by 7 points overall and underestimate by 5 points overall.

This provides a useful way to examine the questions. Students were on average, most adept at estimating their knowledge for the questions with which they had the least familiarity and which proved resilient to test-taking strategies. Test taking strategies were apparent in two specific cases: Although unfamiliar with the content in a question that had as one of its answers "A and B are both correct," many students correctly and confidently selected this answer citing that no other questions had answers containing both. Many students admitted not knowing the answer to a question about the "interaction of several genes" but correctly and confidently chose the answer with the suffix "poly-."

Student Thinking in PFL Scenario.

On questions designed to investigate common misconceptions, students tended to be more strongly overconfident. This does not mean that they are not metacognitive; a student with a persistent misconception may be very confident about the misconception, not knowing it to be a misconception. Similarly, on a question asking about color blindness, a sex-linked trait not covered in class, students were overwhelmingly

overconfident because they treated it as any "normal" trait and felt confident that it would be similar to other traits. Conversely, they dramatically underestimated their understanding for a question on predicting Type O blood. This may be because it is involves more than two alleles, yet is relatively straightforward to solve.

We also asked students to rate their confidence on suggestions made during the scenario. Most students, after searching for information on Sickle Cell Tests, felt confident that they had provided good information. One student reflects: "I am fairly confident in my suggestions and explanations, because thats what i researched and found by researching google." Another student, who was initially quite low in her confidence at the beginning of the scenario became much more confident towards the end:

> I believe that i may have not been accurate with some of my answers but now that i have researched a little more i feel like I know now that most of my answers we correct and i would answer any questions again to make sure that you understand the answers.

Another way to consider the affordances of the PFL assessment over the MC questions is to contrast talk over similar concepts during each type of assessment. Punnett squares are used in genetics to determine the probability that a trait would be inherited based on the traits of the parents. Students who completed the scenario before instruction did not automatically consider a Punnett square, whereas after instruction, most students readily suggested Punnett squares as a means to find the probability of inheriting a trait. However, students were not adept at constructing a Punnett square as previous experience had only involved completing one. Although they learn to associate probability with Punnett squares, the squares themselves have been decontextualized. This is especially apparent in the following conversation,:

> S: So I'm thinking again I can do a Punnett square, I'm not sure. If it was in the answers, it has male and females, so I think --/ R: So that's gonna make it hard for you to do the Punnett square? /S: Yeah. / R:You don't know how to do the Punnett square when there's a male female? /S:Male and female….

Students are not successful at structuring Punnett squares for more than one trait. When using Punnett squares students tend to refer to the traits as letters when talking over the MC questions, but specific traits in the scenario. In the scenario, the Punnett square is contextualized, and each square represents a trait, or more commonly, an individual. While this demonstrates greater agency, it also reveals a common conception students have, relating to probability.

While not all students were asked about probability, of those who were, only two students had robust understanding of chance. Very few students understood that each child would be independent of the next, and that the Punnett square would predict a 25% chance for a certain trait regardless of the traits held by other children. Additionally, some students struggled to predict about the fifth child: "I don't know. I mean, I don't - we only stuck to, like, the four of them."

Another finding related to student conceptions in biology relates to DNA. Although most students know that all cells contain DNA, they confound this with the understanding that cells perform specific functions; they believe cells to contain only the DNA relevant to that particular type of cell. When asked, for instance, if eye color could be determined from a skin cell they explain: "I don't think so. I think they'd be able to tell what color skin and anything that deals with your skin. No. Then they must not have the same DNA."; and

> I wouldn't think they would be able to tell the color of your eyes. Maybe the color of your skin or if you have a skin - they may be able to tell things about your skin...But I don't think as far as your eye color go, they - they probably wouldn't be able to tell.

Because the test for Sickle Cell involves *blood* and the trait results in problems within *blood*, there was nothing in the scenario to push back on this preconception, though we did ask students to explain the cause at the genetic level. However, students commonly explain the cause of the *symptoms* of Sickle Cell Disease. Though they have just answered questions about genetic diseases and searched for information about Sickle Cell Disease, they still focus on superficial aspects: "Well it's nothing to do with their genes or their chromosomes or DNA, nothing like that. It's basically something to do with the blood. I'd tell them that they don't have to worry about their genes and that's it." Another example highlights this common issue:

> S: I don't think there's anything really wrong with their genes, it's just their blood./ R:Okay. So how did they get -- how do they get the disease then? You don't get it through the genes then? It's not a genetic disease, is that what you're saying or --/ S: It's like -- it's a disorder. It's like -- how you get it. Genes is not linked -- let's see. Low oxygen levels increase acidity

> below value [inaudible] of the blood. So it's caused by defect. It's not caused by genes. It's caused by like defects in your blood

Even when students discuss genetic causes, it is clear that this understanding is tenuous: "It is caused by one or both parent's genes, which is what a sickle cell disease is. You can also be born with it or maybe down the line you're not born with it and you could have it then." We did not anticipate that students, especially those from a medical-focused high school, would struggle with the difference between genetic and communicable diseases. We hypothesize that because students primarily reference STD's as examples of communicable diseases, there may be confounding superficial similarities. If your understanding of disease is primarily symptom driven rather than causal, it is likely that a mother passing an STD to her fetus seems the same as passing genetic material.

In the design and piloting of the PFL assessment, we also discovered that scenario based assessments such as these allow the students and teachers to take learning in many directions, a possibility not afforded by multiple choice tests. For example, bioethics is a topic often discussed in high school biology curricula, but it is difficult to gauge whether students have come to have deep ethical understanding of such a complex topic. This scenario provided information that would allow an instructor to see that a student is operating under a false pretense either related to subject matter as we have demonstrated, or in this case, about the role of a professional, the genetic counselor. Genetic counselors provide explanations and options for couples, but do not tell couples what to decide. This is stated in the scenario, yet clearly not taken up by all students:

> S: I think they should adopt. R: You think they should adopt? S: Cause they'll have to spend all their money if the child is sick without the insurance and the hospital bills and stuff. And I know they don't want to be like worried that their child gonna die from sickle cell, so… R: And why do you think that? S: Because adoption, I think that's the best option.

Additionally, we have found evidence of learning. . In one case, a student was initially confident that the couple would not be at risk for passing on Sickle Cell:

> I think -- I'm thinking they should have a child because looking at this information here, after her grandparent had it, but they did not pass it on to any of their children and his great-grandparents had it, but didn't pass it on to any of their children. So I think they're fine. They can have children without them being able to pass this sickle cell disease to them.

After researching, she discovers that carriers are generally healthy and that if both parents are carriers, there is a risk:

> S: so this means that if they do have a child that it would possibly have sickle cell./ R: Okay. And how do you know that now? You've changed your mind about that. /S: Because now that I have the blood test that sees that they both have the trait. /R: Okay. /S: Of what I had read, it said that both the mother and father had to have possessed the trait for them to carry sickle cell, so now that I know they both carry the trait, if they have a child, they will pass it on. /R: Okay. So they definitely -- their child will have it? /S: Well, no, I don't think -- I won't say definitely they will, but it's a good possibility they will. [….] S: I would like them to know the symptoms of sickle cell and what different things they will have to look for in disease. And I would -- when -- if they do have it, if they do decide to have their child, I would like -- I would say they need to test the child.

There are two other general findings that we would like to point out. The first relates to students who were reluctant to use the internet to search for answers to some of the questions contained in the scenario: This is relevant as it shows that not all students are engaging in skills that are required for the 21st century. If this type of assessment is used, the students will, by default, be required to take part in activities that better prepare them for the future.

An exciting finding that we had hopefully anticipated is that the PFL assessment is indeed the instructional tool that we intended it to be. Following many of the sessions, researchers had the opportunity to informally discuss with subjects their impressions of the test. We can infer that students felt that the tool helped them learn:

> R: So, how familiar were you with sickle cell disease before this session? S: I knew a little bit about, but I didn't know that much about it R: Did you learn something about it? S: Um, yes. [inaudible] when parents can have [inaudible] self.

Discussion

The development of an assessment tool that is both scalable and capable of making student thinking and learning more visible presents a possible solution to a critical global need. While there are examples of assessments making use of computers, and even use of latent semantic analysis for evaluating essays, neither leverages the potential presented by modern technology. As we iteratively develop our PFL assessment, we tend to agree with Schank (1996) that new cases should violate expectations from prior cases, such that they challenge preconceptions and allow students to apply their learning across contexts. We intend to develop cases that ask the student-as-genetic counselor to consider ethical issues, as well as to assist geneticists in the design of research studies. Simulated experts, rather than providing their expertise (Bell, Bareiss, & Beckwith, 1993) will provide further questions, pushing students to critically examine their understanding. Instead of allowing students to choose an option out of a few choices (Bell, Bareiss, & Beckwith, 1993), (and mimicking the very thing we are striving to depart from), we intend for the student to direct his or her own learning and research, interspaced with formative feedback in the form of questions from "colleagues."

Though our current pilot of the assessment does not yet do this, one of our goals is to understand how to incorporate technology in impactful ways. In the current version, web-based searching is included as a way to both understand what skills and experience students bring to the task, and to consider how we can enable them to move beyond their current states *during* the assessment. We recognize the shortcomings of a largely text-based assessment of scientific skills and intend for future renditions to foster model-based thinking (Schauble, 1996; Schauble, Glaser, Raghavan, & Reiner, 1991) as well as encourage students to engage in complex ways with dynamic systems.

Implications and Future Directions

We have presented findings from a design-based research pilot study investigating the affordances of PFL Assessments. We are iterating towards a scalable "working smart" assessment system that will show how students—individually and in teams—learn and improve. The assessment system will use problem-based curricula in which students participate in cycles of 1) self-, technology- and teacher-directed learning, 2) formative assessment of understanding and learning processes, 3) further learning and revision, and 4) benchmark assessment of knowledge and skills. To this end, we will iteratively redesign PFL scenarios and critically examine their small-scale integration into classrooms, particularly contrasting individual and group participation. We will consider opportunities for formative feedback and investigate technologies to aid in real-time analysis of student responses. As we prepare to scale up, we must examine implementations in diverse test-beds, and consider the affordances of technology for embedding working smart tools for both formative and summative assessments. In this process, we seek to continuously improve and to find partners who can contribute to this substantial undertaking.

References

Anderson, L. W., Sosniak, L. A., & Bloom, B. S. (1994). *Bloom's taxonomy: a forty-year retrospective*. Yearbook of the National Society for the Study of Education, 93rd, pt. 2. Chicago: NSSE.

Au, W. (2007). High-Stakes Testing and Curricular Control: A Qualitative Metasynthesis. *Educational Researcher, 36*(5), 258.

Barron, B., Schwartz, D. L., Vye, N. J., Moore, A., Petrosino, A. J., Zech, L., et al. (1998). Doing with Understanding: Lessons from Research on Problem-and Project-Based Learning. *The Journal of the Learning Sciences, 7*(3/4), 271-311.

Bell, B., Bareiss, R., & Beckwith, R. (1993). Sickle Cell Counselor: A Prototype Goal-Based Scenario for Instruction in a Museum Environment. *Journal of the Learning Sciences, 3*(4), 347-386.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education, 5*(1), 7-74.

Bond, L. A. (1996). Norm- and criterion-referenced testing. *Practical Assessment, Research & Evaluation* 5(2). Retrieved November 17, 2007, from http://PAREonline.net/getvn.asp?v=5&n=2

Bransford, J. D., Brown, A. L., & Cocking, R. R. (2000). *How People Learn: Brain, Mind, Experience, and School. Expanded Edition*.

Bransford, J. D., & Schwartz, D. L. (1999). Rethinking Transfer: A Simple Proposal with Multiple Implications. *Review of Research in Education, 24*, 61-100.

Glaser, B. G., & Strauss, A. L. (1967). *The Discovery of Grounded Theory: Strategies for Qualitative Research*: Aldine Transaction.

Means, B. (2006). Prospects for Transforming Schools with Technology-Supported Assessment. In R. K. Sawyer (Ed.), *The Cambridge Handbook of the Learning Sciences*. Cambridge: Cambridge University Press.

Mislevy, R. J., & Riconscente, M. M. (2005). *Evidence-Centered Assessment Design: Layers, Structures, and Terminology (PADI Technical Report 9)*. Menlo Park, CA: SRI International.

Partnership for 21st Century Skills. (2003). *Learning for the 21st Century: A Report and MILE Guide for 21st Century Skills*. Washington, DC: Partnership for 21st Century Skills.
Popham, W. J. (2000). Big Change Questions." Should Large-Scle Assessment be used for Accountability?"--Answer: Depends on the Assessment, Silly! *Journal of Educational Change, 1*(3), 283-289.
Quellmalz, E., DeBarger, A., Haertel, G., Schank, P., Buckley, B., Gobert, J., Horwitz, P., & Ayala, C. (2007). *Exploring the Role of Technology-Based Simulations in Science Assessment: Calipers Project*. Paper presented at the Conference Name|. Retrieved Access Date|. from URL|.
Schank, R. C. (1996). Goal-Based Scenarios: Case-Based Reasoning Meets Learning by Doing. In D. Leake (Ed.), *Case-Based Reasoning: Experiences, Lessons & Future Directions* (pp. 295-347): AAAI Press/The MIT Press.
Schauble, L. (1996). The development of scientific reasoning in knowledge-rich contexts. *Developmental Psychology, 32*(1), 102-119.
Schauble, L., Glaser, R., Raghavan, K., & Reiner, M. (1991). Causal Models and Experimentation Strategies in Scientific Reasoning. *The Journal of the Learning Sciences, 1*(2), 201-238.
Wilson, M., & Sloane, K. (2000). From Principles to Practice: An Embedded Assessment System. *Applied Measurement in Education, 13*(2), 181-208.
Wilson, S. M. (2003). *California Dreaming: Reforming Mathematics Education*: Yale University Press.
Wineburg, S. (2004). Crazy for History. *The Journal of American History, 90*, 1401-1414.

Acknowledgments

We would like to thank the many people of North Carolina who have supported this project. Sam Houston, President and CEO of the North Carolina Science, Mathematics and Technology Center was particularly instrumental in supporting this project and funding this project through the Burroughs-Wellcome Fund. Preparation of this paper was strongly influenced by our colleagues, and supported in part through an NSF grant # 0354453. However, the ideas expressed in this article are not necessarily those of the NSF. Although we are unable to cite them by name to protect their identity, we would especially like to thank the students, teachers, faculty and staff at the school where we used the instructional tool. They were extremely generous, thoughtful and committed.