Middle School Students' Use of Appropriate and Inappropriate Evidence in Writing Scientific Explanations

Katherine L. McNeill & Joseph Krajcik

University of Michigan

contact info:
Center for Highly Interactive Computing in Education
610 E. University Ave., Ann Arbor, MI, 48109-1259
734-647-4226
kmcneill@umich.edu

Reference as:
McNeill, K. L. & Krajcik, J. (in press). Middle school students' use of appropriate and inappropriate evidence in writing scientific explanations. In Lovett, M & Shah, P (Eds.) *Thinking with Data: the Proceedings of the 33rd Carnegie Symposium on Cognition.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Abstract

Recent science reform efforts and standards documents advocate that students develop scientific inquiry practices, such as the construction and communication of scientific explanations. This paper focuses on 7th grade students' scientific explanations during the enactment of a project based chemistry unit where the construction of scientific explanations is a key learning goal. During the unit, we make the explanation framework explicit to students and include supports or scaffolds in both the student and teacher materials to facilitate students' in their understanding and construction of scientific explanations. Results from the enactment show significant learning gains for students for all components of scientific explanation (i.e. claim, evidence, and reasoning). Although students' explanations were stronger at the end of the instructional unit, we also found that students' still had difficulty differentiating between appropriate and inappropriate evidence for some assessment tasks. We conjecture that students' ability to use appropriate data as evidence depends on the wording of the assessment task, students' content knowledge, and their understanding of what counts as evidence. Having students construct scientific explanations can be an important tool to help make students thinking visible for both researchers and teachers.

Middle School Students' Use of Appropriate and Inappropriate Evidence in Writing Scientific Explanations

The National Research Council (1996) and the American Association for the Advancement of Science (1993) call for scientific literacy for all. All students need knowledge of scientific concepts and inquiry practices required for personal decision making, participation in societal and cultural affairs, and economic productivity. Science education should support students' development toward competent participation in a science infused world (McGinn & Roth, 1999). This type of participation should be obtainable for all students, not just those who are educated for scientific professions. Consequently, we are interested in supporting all students in learning scientific concepts and inquiry practices.

By scientific inquiry practices, we mean the multiple ways of knowing which scientists use to study the natural world (National Research Council, 1996). Key scientific inquiry practices called for by national standards documents include asking questions, designing experiments, analyzing data, and constructing explanations (American Association for the Advancement of Science, 1993; National Research Council, 1996). In this study, we focus on analyzing data and constructing explanations. These practices are essential not only for scientists, but for all individuals. On a daily basis, individuals need to evaluate scientific data provided to them in written form such as newspapers and magazines as well spoken through television and radio. Citizens need to be able to evaluate that data to determine whether the claims being made based on the data and reasoning are valid. This type of data evaluation, like other scientific inquiry practices, is dependent both on a general understanding of how to evaluate data as well as an understanding of the science content.

In this study we explore when students use appropriate evidence and when they use inappropriate evidence to support their claims. Our work focuses on an 8-week project-based chemistry curriculum designed to support 7th grade students in using evidence and constructing scientific explanations. We examine the characteristics of these students' explanations, their understanding of the content knowledge, and the assessment tasks to unpack what may be influencing students use of evidence.

Our Instructional Model for Scientific Explanations

In our work, we examine how students construct scientific explanations using evidence. We use a specific instructional model for evidence-based scientific explanations as a tool for both classroom practice and research. We provide both teachers and students with this model to make the typically implicit framework of explanation, explicit to both teachers and students.

Our instructional model for scientific explanation uses an adapted version of Toulmin's (1958) model of argumentation and builds off previous science educators' research on students' construction of scientific explanations and arguments (Bell & Linn, 2000; Jiménez-Aleixandre, Rodríguez, & Duschl, 2000; Lee & Songer, 2004; Sandoval, 2003; Zembal-Saul, et al., 2002). Our explanation framework includes three components: a claim (similar to Toulmin's claim), evidence (similar to Toulmin's data), and reasoning (a combination of Toulmin's warrants and backing). The claim makes an assertion or conclusion that addresses the original question or problem. The evidence supports the student's claim using scientific data. This data can come from an investigation that students complete or from another source, such as observations, reading material, or archived data. The data need to be both appropriate and sufficient to support the claim. Appropriate data is relevant to the question or problem and relates to the given claim. Data is sufficient when it includes the necessary quantity to convince someone of a claim. The

reasoning is a justification that links the claim and evidence and shows why the data counts as evidence to support the claim by using the appropriate scientific principles.

Kuhn argues (1993) that argument, or in our case scientific explanation, is a form of thinking that transcends the particular content to which it refers. Students can construct scientific explanations across different content areas. Although an explanation model, such as Toulmin's, can be used to assess the structure of an explanation, it cannot determine the scientific accuracy of the explanation (Driver, Newton & Osborne, 2000). Instead, both the domain general explanation framework and the domain specific context of the assessment task determine the correctness of the explanation. Consequently, in both teaching students about explanation and assessing students' construction of explanations we embed the scientific inquiry practice in a specific context.

Student Difficulties Constructing Explanations

Prior research in science classrooms suggests that students have difficulty constructing high-quality scientific explanations where they articulate and defend their claims (Sadler, 2004). For example, students have difficulty understanding what counts as evidence (Sadler, 2004) and using appropriate evidence (Sandoval, 2003; Sandoval & Reiser, 1997). Instead, students will draw on data that do not support their claim. Consequently, we are interested in whether students use appropriate evidence to support their claim or if they draw on evidence that is not relevant.

Students' claims also do not necessarily relate to their evidence. Instead, students often rely on their personal views instead of evidence to draw conclusions (Hogan & Maglienti, 2001). Students have a particularly difficult time reasoning from primary data, especially when

measurement error plays an important role (Kanari & Millar, 2004). Students can recognize variation in data and use characteristics of data in their reasoning, but their ability to draw final conclusions from that data can depend on the context. Masnick, Klahr, and Morris (this volume) concluded that young students who poorly understood the context of the investigation had difficulty interpreting data, particularly when the interpretation of that data contradicted their prior beliefs. Students will likely discount data if the data contradicts their current theory (Chinn & Brewer, 2001) and they will only consider data if they can come up with a mechanism for the pattern of data (Koslowski, 1996). When students evaluate data, more general reasoning strategies interact with domain-specific knowledge (Chinn & Brewer, 2001). Whether students use appropriate and inappropriate evidence may depend on their prior understanding of a particular content area or task.

Students also have difficulty providing the backing, or what we refer to as reasoning, for why they chose the evidence (Bell & Linn, 2000) in their written explanations. Other researchers have shown that during classroom discourse, discussions tend to be dominated by claims with little backing to support their claims (Jiménez-Aleixandre, Rodríguez & Duschl, 2000). Our previous work supports these ideas. We found that middle school students' had the most difficulty with the reasoning component of scientific explanations (McNeill, Lizotte, Krajcik & Marx, in review; McNeill, et al., 2003). Although students' reasoning improved over the course of the 6-8 week instructional unit, it was consistently of lower quality than their claims or evidence. Students' reasoning often just linked their claim and evidence and less frequently articulated the scientific principles that allowed them to make that connection.

Similar to students ability to evaluate and use data, providing accurate reasoning is related to students understanding of the content. Students with stronger content knowledge

provide stronger reasoning in their scientific explanations (McNeill et al., in review). Previous research with students has found that their success at completing scientific inquiry practices is highly dependent on their understanding of both the content and the scientific inquiry practices (Metz, 2000). Both domain specific and general reasoning are essential for students' effective evaluation of data and construction of scientific explanations.

Although previous work has shown that students have difficulty with components of scientific explanations, there has been little research unpacking exactly when students have difficulty or why they have difficulty. In this paper, we address the following research question: What difficulties do middle school students have using evidence and reasoning when constructing scientific explanations? How does content area influence students use of evidence and reasoning? Furthermore, we attempt to unpack what may be causing these student difficulties.

Instructional Context

Our Model of Learning

Our work is rooted in social constructivist learning theories that argue that understanding is contextualized and a function of social interactions with others (Blumenfeld et al., 1997; Driver et al., 1994; Singer et al., 2000). Our model of learning stems from five important features: *active construction*, *situated cognition*, *community, discourse,* and *cognitive tools* (Rivet & Krajcik, 2004; Singer et. al, 2000). *Active construction* of knowledge states that students create new knowledge and understanding based on what they already know and believe. This knowledge includes not only content knowledge, but also knowledge students have acquired because of their social roles connected with race, class, gender and their cultural and ethnic affiliations (Bransford, Brown & Cocking, 2000). *Situated cognition* recognizes that

learning is a social process and students make meaning through their interactions with other people, tools, and the environment (Lave & Wenger, 1991). Within the classroom, these interactions occur in a *community* of practice where students learn to learn from their teacher, peers, and other resources (Brown et al., 1993). Science has its own special *discourse* and that language is the primary means for communicating scientific knowledge (Lemke, 1990). Students need to learn how to talk using scientific discourse and not just talk about science. Scientific discourse does not just mean scientific words, but it includes different ways of knowing in science, such as asking questions, designing experiments, and constructing explanations (Driver et al, 1994). Another important component of the classroom environment is *cognitive tools*. Cognitive tools provide supporting structures to students that act as intellectual partners to extend performance and learning (Solomon, Perkins & Globerson, 1991). Cognitive tools can range from computer software that allows students to create complex representations to written instructional scaffolds that encourage students to include evidence in their explanations. In all cases, they allow students to function at a level beyond their own independent cognitive abilities to engage in complex problem solving.

Learning-Goals-Driven Design Model

To engage middle school students in developing a deep understanding of science content and scientific inquiry practices, we developed a 7th grade chemistry unit, "How can I make new stuff from old stuff?" (referred to as "Stuff") that sustains students in learning over a six to eight week period (McNeill, Harris, Heitzman, Lizotte, Sutherland, & Krajcik, 2004). We designed the *Stuff* unit using a *learning-goals-driven design* process, based upon the backwards design model of Wiggins and McTighe (1998). The central focus is to identify learning goals and derive

learning performances that illustrate how students should use the scientific content and practices in real tasks (Reiser et al., 2003). *Learning performances* reflect the reasoning tasks we want students to be able to do with scientific knowledge. Learning performances reformulate a scientific content standard in terms of scientific practices that *use* that content, such as students being able to define terms, describe phenomena, use models to explain patterns in data, construct scientific explanations, or test hypotheses. The articulated learning performances serve as guides for designing activities and assessments. We developed learning performance by crossing a content specific standard with a scientific inquiry practice standard. Table 1, gives an example.

INSERT TABLE 1

Description of Stuff Unit

Stuff engages students in the study of substances and properties, the nature of chemical reactions, and the conservation of matter. In the *Stuff* unit, we contextualized the concepts and scientific inquiry in real world experience by focusing on making soap from fat or lard and sodium hydroxide (making new stuff from old stuff). Students complete a number of investigations where they revisit soap and fat throughout the unit. These cycles help students delve deeper into the key learning goals including both target science content and the scientific inquiry practices such as the analysis of data and construction of scientific explanations.

Initially, students explore two unknowns (soap and fat) to introduce the concepts of substance and property. In order to develop students' understanding of properties, they investigate solubility, melting point, and density for soap and fat. Next, students explore a number of chemical reactions by observing macroscopic phenomena, including making soap from fat. Finally, students' build on their understandings by exploring what happens to mass

during chemical reactions. Throughout the unit students alternate between exploring the macroscopic phenomena and using molecular models to explain the phenomena. The unit ends with testing the properties of their homemade soap to determine whether or not they created a new substance.

Supporting Students' Understanding of Scientific Explanations

To support students' construction of scientific explanation we embedded several strategies into our instructional unit including: making the rationale behind explanation explicit, modeling how to construct explanations, providing students with opportunities to engage in explanation construction, and including written scaffolds on students' investigation sheets.

Revealing the tacit framework of scientific explanation can facilitate students' explanation construction (Reiser et al, 2001). We accomplish this through both written scaffolds in the instructional sheets and encouraging teacher practices that help students understand this framework. Initially, teachers introduce the framework for scientific explanation during a focal lesson where they discuss the importance of explanation construction and define the three components of scientific explanation. In our previous work, we have found that when teachers discuss the rationale behind scientific explanations, that students construct stronger explanations (Lizotte, McNeill, & Krajcik, 2004).

Also, during the focal lesson teachers model how to construct an explanation. Teacher modeling of scientific inquiry practices can result in more effective learning environments (Crawford, 2000; Tabak & Reiser, 1997). Specifically, for scientific explanation modeling can help students engage in this practice (Lizotte, McNeill, & Krajcik, 2004). After the initial focal lesson, we encourage teachers to continue modeling explanation construction throughout the

unit. The student readers also include both strong and weak examples of scientific explanations for the teacher and students to critique.

In order for students to learn how to evaluate data, they need numerous opportunities to evaluate rich, complex models of data (Chinn & Brewer, 2001; Lehrer & Schauble, 2002). We assume that students also need numerous opportunities to engage in scientific explanations. Over the course of the unit, students construct at least ten explanations.

Written scaffolds embedded in student materials can scaffold students' development of scientific inquiry practices, modeling, and metacognitive skills and knowledge (White & Frederiksen, 1998). Our work builds off of this research on written scaffolds as well as our previous research where we found that fading written scaffolds resulted in students constructing stronger explanations (McNeill, et al., in review). We provide scaffolds in the student investigation sheets that initially define the three components of scientific explanations (i.e. claim, evidence, and reasoning) and then fade over time.

Method

Participants

For this study, we report findings from teachers in urban and suburban sites in the Midwest that enacted the *Stuff* unit. Teachers and students in the urban site came from a large urban district implementing reform-based curricula. The teachers and students were predominately African American with the students coming from lower to lower-middle income families. The teachers and students from the suburban site were from an independent middle school in a large college town also involved in implementing reform-based curricula. The teachers and the majority of the students were Caucasian with the students coming from middle

to upper-middle income families. It was the second time for three teachers in the urban school and for the three teachers in the suburban site enacting the *Stuff* unit. All teachers completed lessons necessary for the study, with the time of enactment ranging from 5 1/2 weeks to 8 weeks. Table 2 shows the breakdown of teachers, classroom and students in the two sites for the 2003-2004 school year.

INSERT TABLE 2

Assessment Data

All students completed identical pre and posttest measures that included 15 multiple-choice items and 4 open-ended responses. Only students who completed all parts of the test were included in the analysis. Due to high absenteeism and mobility in the urban schools, a number of students did not complete both the pre and posttests. Consequently, our analysis only includes 700 of the students.

Multiple-choice responses were scored and tallied for a maximum possible score of 15. We developed rubrics to score the 4 open-ended items. Maximum score on the open-ended items was 15. All questions were scored by one rater. We then randomly sampled 20% of the open-ended test items and a second independent rater scored them. For each of the four open-ended test items our estimates of inter-rater reliability were calculated by percent agreements. Our inter-rater agreement was above 93% for each component (i.e. claim, evidence, and reasoning) of each question.

The multiple-choice items covered the three key learning goals of the unit: substance and properties, chemical reactions, and conservation of mass. Appendix A includes four sample multiple-choice items that align with the substance and properties and chemical reaction learning

goals, which are the focus of this analysis. Appendix B includes the open-ended items that asked students to write scientific explanations for these two different content areas. Both items include appropriate evidence (e.g. density and melting point) that should be used to answer the question and inappropriate evidence (e.g. mass and volume) that is not relevant to the particular task.

To assess student understanding of scientific explanation, we developed a *base* rubric to use across different content areas (Harris, et al. in press). We used our base rubrics to develop *specific* rubrics for assessing students on each learning and assessment task for our chemistry unit. Appendix C includes the specific rubrics we used to score the two explanation tasks on the pre and posttest. The rubric includes the three components of scientific explanation (claim, evidence, and reasoning) and discusses the criteria for different levels of each component. For example, for the highest score for evidence students need to include appropriate evidence that addresses the particular task, sufficient evidence in that they include enough data to support the claim, and not include any inappropriate evidence. We calculated students' total evidence scores by subtracting the number of appropriate pieces evidence minus one if they included any inappropriate evidence.[1]

We discuss three examples from the substance and property explanation to demonstrate how we used the rubrics to score students' responses. Student A provides an example of a strong explanation.[2]

---

[1] However, we did not give students negative evidence scores. If they received a zero for appropriate evidence and a one for inappropriate evidence, there total evidence score was still recorded as a zero. Yet we kept track that these students had received a one for inappropriate evidence for other analyses.

[2] Students' original spelling, grammar, and punctuation are intact in this and all future examples.

> Student A:
> Liquid 1 and 4 are indeed the same substance. Looking at this data, the properties include density, color, and melting point. Mass is not a property. Density, color, and M.P. are all the same for liquid 1 and 4. Since all of these properties are the same, 1 and 4 are the same substance.

This student provides an accurate claim that liquid 1 and 4 are the same substance. She provides three pieces of appropriate evidence (density, color and melting point) and no inappropriate evidence. She also includes the highest level of reasoning because she states that the same substances have the same properties.

The second example, Student B, provides an example of a weak explanation.

> Student B:
> No, the liquids are not the same substance because some are different and some are the same like Liquid 1 and Liquid 4 is just that liquid 4 has different mass than liquid 1. Some has color like Liquid 1, Liquid 2, and liquid 4 they all has no color but liquid 3 the color is silver. And also liquid 2, and liquid 3 has different density than liquid 1 & 2. Liquids 2 & 3 has different melting point.

This student received zeros for claim, appropriate evidence and reasoning. The student received a score for inappropriate evidence because he included mass as data for determining whether two liquids are the same substance.

The final example, Student C, provides a mixed response. The response includes both correct and incorrect components.

> Student C:
> Out of the 4 liquids, none of them are alike. They aren't alike because the density of 2 & 3 are different from 1 & 4. The color of liquid 3 is different from 2, 1, & 4. The mass of 3 & 4 is different from 1 & 2. The melting point of 1 & 4 is different from 2 & 3. In order for these liquids to be the same substances, they must have the same properties.

This student received a zero for claim and evidence, because she did not provide the claim that Liquids 1 and 4 are the same substance or provide any evidence to support why the substances are the same. Student C received a score for inappropriate evidence, because she included mass

as important for determining whether two liquids are the same substance. Although the student receives low scores for claim and evidence, she did receive a high score for reasoning. The last sentence includes a correct scientific statement that in order for two substances to be the same, they must have the same properties. This example illustrates how the rubric codes each component independent of the score received on the other two components. Even though Student C was unable to construct an accurate claim or evidence, she did receive credit in her reasoning for having some understanding of the underlying general scientific principle. Consequently, the rubric allows us to tease apart a student's understanding of a particular component, though a drawback is that it does not provide a holistic score of the overall coherence of the explanation.

Results and Discussion

In this study we examine when students use appropriate evidence and when they use inappropriate evidence to support their claims. To address this overarching question, our analyses address the following sub-questions: 1) Do students achieve learning gains for both the science content and the different components of scientific explanations during the unit? 2) What difficulties do middle school students have using appropriate evidence when constructing scientific explanations? 3) How does the content area and task influence students' use of appropriate and inappropriate evidence? We then explore possible causes for students' use of appropriate and inappropriate evidence.

Student Learning Gains

Before examining students' learning for scientific explanations, we first examined

whether students learned the key learning goals addressed in the unit. Figure 1 provides the student mean scores for the multiple-choice items, the open ended questions, and the total test score for the pre and posttests. We conducted one-tailed paired *t*-tests to test the significance of students' learning gains. Students achieved significant learning gains on the multiple choice, $t(699) = 37.19$, $p < .001$, open-ended, $t(699) = 33.96$, $p < .001$, and total test score, $t(699) = 42.85$, $p < .001$. The effects sizes for student learning for the multiple-choice, open-ended and total scores were 1.81, 2.05 and 2.34 respectively[3]. This suggests that students had a much stronger understanding of the content and scientific inquiry practices after the instructional unit.

INSERT FIGURE 1

Next, we examined students' learning for the evidence-based scientific explanations they constructed. Figure 2 provides the means for student scores for the different components of scientific explanation. Again, we conducted one-tailed paired *t*-tests to test the significance of students' learning gains. Students achieved significant learning gains on claim, $t(699) = 23.93$, $p < .001$, evidence, $t(699) = 22.43$, $p < .001$, and reasoning, $t(699) = 25.77$, $p < .001$.

INSERT FIGURE 2

The effect sizes were 1.24 for claim, 1.51 for evidence, and 3.75 for reasoning. Interestingly, in the previous two enactments we found that students' reasoning was consistently much lower than their claims and evidence (McNeill, et al., in review; McNeill, et al., 2003).

[3] Effect Size was calculated by dividing the difference between posttest and pretest mean scores by the pretest standard deviation.

Consequently, we made revisions to the instructional unit and professional development to help both teachers and students with the reasoning component. The results from the present study show that students' reasoning scores had greater learning gains and were closer to their evidence score by the end of the unit than in previous studies. However, students' evidence and reasoning were still lower than their claim scores and low at an absolute level. These results suggest that providing evidence and reasoning continued to challenge students. Consequently, in this study we further unpack potential causes of these difficulties.

Students' Explanation Scores By Content Area

We examined whether students' learning gains and overall performance for scientific explanations differed by content area. Figures 3 and 4 provide the breakdown for claim, evidence, and reasoning for two content areas: substance/property and chemical reactions. For both explanations, students achieved significant learning gains for claim, evidence, and reasoning (ps < .001).

INSERT FIGURE 3 & FIGURE 4

Comparing the posttest values for the two explanations shows that students scored higher on the chemical reaction claim than the substance and property claim[4]. Yet they scored lower on the chemical reaction evidence and reasoning. We found these results surprising. We would have predicted that if students scored higher on claim, they would also score higher on evidence and reasoning because even though we coded each component independently students use the

[4] We weighted students claim, evidence, and reasoning scores so that the maximum score was 1.25 for each component for both the substance and property explanation and the chemical reaction explanation.

evidence and reasoning to construct their claims. To further unpack this trend, we examined students' use of both appropriate and inappropriate evidence.

Students' Use of Inappropriate Evidence

We explored whether students used inappropriate data in their explanations and whether this differed for the two content areas. In examining students' responses, we found that of the 700 students, 125 provided inappropriate evidence for the substance and property explanation, while 184 students provided inappropriate evidence for the chemical reaction explanation (see Table 3). This suggests that students were more likely to include inappropriate evidence for the chemical reaction explanation. This is one possible reason for why students' evidence scores were lower for the chemical reaction explanation than the substance and property explanation. The lower total evidence scores might have been a result of this greater use of inappropriate evidence, since we calculated the total evidence score by subtracting the inappropriate evidence from the appropriate evidence.

INSERT TABLE 3

Table 3 shows that 43 students provided inappropriate evidence for both explanations. Overall the majority of students included inappropriate evidence in only one question or the other. For the substance and property explanation 80 students provided inappropriate evidence who do not for the chemical reaction question, while for the chemical reaction question 141 students provided inappropriate evidence who do not on the substance and property question.

Consequently, we examined, which students provided inappropriate evidence and which

students provided appropriate evidence for the two questions. By exploring students' use of evidence, we hoped to come up with some initial hypothesis of why students provided inappropriate evidence. Such findings would allow us to provide guidance to the field in how to help students provide appropriate evidence.

Substance and property explanation. The substance and property explanation item (Appendix B) includes three pieces of appropriate evidence (density, color, and melting point) and one piece of inappropriate evidence (mass). We were interested in how similar and different students were who included the inappropriate evidence (i.e. mass) in their responses compared to those who did not include inappropriate evidence in terms of their claims, reasoning, and content knowledge. We predicted that students who did not include inappropriate evidence would have stronger claims, reasoning, and content knowledge.

To test whether the students differed, we conducted a two-way analysis of variance (ANOVA) where we split students into four groups based on their use of evidence on the posttest for the substance and property explanation. The four groups consisted of students who used: 1. No appropriate evidence and inappropriate evidence, 2. No appropriate evidence and no inappropriate evidence, 3. Appropriate evidence and inappropriate evidence, and 4. Appropriate evidence and no inappropriate evidence. The results from the two-way ANOVA for students' posttest claim and reasoning scores by their use of appropriate and inappropriate evidence are shown in Figure 5. There is a significant difference in the scores for the four groups of students for claim, F (3, 696) = 191.287, $p < .001$, and reasoning, F (3, 696) = 5.991, $p < .001$.

INSERT FIGURE 5

Students who used appropriate evidence (i.e. density, color and melting point), but did not use inappropriate evidence (i.e. mass) had the highest claim and reasoning scores. This matches our predictions of what we thought would occur. Even though the rubric scores each component independent of the others, students typically base their claims on their evidence and reasoning. Consequently, we would expect that students with higher claims would also have higher evidence and reasoning scores. For the substance and property item, the use of appropriate evidence appears to be particularly important for constructing the valid claim that liquids 1 and 4 are the same substance. The two groups of students that used appropriate evidence scored higher than the two groups that did not use appropriate evidence.

We also explored if a relationship existed between students' understanding of the content and their use of appropriate and inappropriate evidence. We used students' scores on the multiple-choice items on the posttest for the substance and property items as a measure of their understanding of the content. To test whether the four groups of students differed in their understanding of the content, we completed a two-way ANOVA for students' multiple-choice scores by their use of appropriate and inappropriate evidence. Figure 6 displays the results from this analysis. A significant difference exists for the four groups of students' multiple-choice scores, F (3, 696) = 12.947, $p < .001$. Students who used appropriate evidence, but did not use inappropriate evidence had the highest content score. This suggests a relationship between students' understanding of the content and their ability to use appropriate evidence. Overall, students' scores on the substance/property multiple-choice items were significantly correlated with their substance/property explanations, rs (700) = 0.26 for claim, 0.23 for evidence, and 0.35 for reasoning, ps < .001. Students who had higher multiple-choice scores in a content area also had higher explanation scores in that area.

INSERT FIGURE 6

Students who have a stronger understanding of the content are more likely to include appropriate evidence and less likely to include inappropriate evidence. Furthermore, students who include appropriate evidence and do not include inappropriate evidence are more likely to construct stronger claims and reasoning.

Chemical reaction explanation. We expected to find similar results to the substance and property explanation for the chemical reaction explanation. We predicted that students who did include appropriate evidence and did not include inappropriate evidence would have stronger claims, reasoning, and content knowledge. In this explanation item (Appendix B) there are three pieces of appropriate evidence (density, melting point, and solubility) and two pieces of inappropriate evidence (mass and volume). We categorized students as including inappropriate evidence if they used either or both mass and volume.

Again, we split students into four groups based on their use of evidence: 1. No appropriate evidence and inappropriate evidence, 2. No appropriate evidence and no inappropriate evidence, 3. Appropriate evidence and inappropriate evidence, and 4. Appropriate evidence and no inappropriate evidence. To test whether the four groups differed we completed a two-way ANOVA for students' posttest claim and reasoning scores by their use of appropriate and inappropriate evidence. A significant difference in the scores for the four groups of students exists for claim, $F\,(3, 696) = 42.979$, $p < .001$, and reasoning, $F\,(3, 696) = 7.311$, $p < .001$ (Figure 7).

INSERT FIGURE 7

The trend for students' claim scores is different than in the substance and property explanation. Although students who included no appropriate and no inappropriate evidence had lower claim scores, the claim scores for the other three groups of students did not differ. Particularly noteworthy is that students who included inappropriate evidence, but no appropriate evidence had similar claim scores to students who included appropriate evidence, but no inappropriate evidence. This suggests that students were able to use inappropriate evidence (i.e. mass and weight) to come up with the correct claim that a chemical reaction did occur. Students reasoning scores were similar to the substance and property explanation. Students who used appropriate evidence, but no inappropriate evidence again provided the strongest reasoning.

We also tested whether the four groups of students differed in their understanding of the content by completing a two-way ANOVA for students' multiple-choice scores by their use of appropriate and inappropriate evidence. Figure 8 displays the results from this analysis. There is a significant difference, F (3, 696) = 29.335, $p < .001$. Similar to the substance and property explanation, students who used appropriate evidence, but who did not use inappropriate evidence had higher content knowledge. Again, we also see that students' scores on the chemical reaction multiple-choice items were significantly correlated with their chemical reaction explanations, rs (700) = 0.23 for claim, 0.33 for evidence, and 0.29 for reasoning, ps < .001.

INSERT FIGURE 8

The reasoning and content analysis showed similar trends across the two explanations, but students' use of appropriate and inappropriate evidence to support their claims varied. Although a relationship existed between using appropriate evidence and creating the correct claim for the substance and property explanation, that same relationship did not exist for the chemical reaction explanation. For the chemical reaction explanation, students who provided no evidence had lower claim scores, yet students who used inappropriate evidence were just as likely to construct the correct claim as those who use appropriate evidence. In order to investigate what might have caused these students to use inappropriate evidence in the chemical reaction explanation, we reexamined the assessment items and examples of student work. This analysis also provides possible reasons for why overall students have higher claims, yet lower evidence and reasoning scores for the chemical reaction explanation compared to the substance and property explanation.

Exploration of Why Students Used Inappropriate Evidence

Differences in the wording of the assessment task. First we examined why students who used inappropriate evidence were less likely to make the correct claim for the substance property explanation yet more likely to make the correct claim for the chemical reaction explanation. Looking back at these two explanation questions (Appendix B), we realized that for the chemical reaction question students can actually use the inappropriate evidence to make the correct claim. Students can examine the question and see that the mass and volume changed. Consequently, they can claim that a chemical reaction occurred because the mass and/or volume changed from before stirring and heating to after stirring and heating. In this case, we gave them credit for providing the correct claim even though they used incorrect evidence to get there. The student

response below is one example of a student who used inappropriate evidence to construct an accurate claim for the chemical reaction explanation.

> Student D:
> A chemical reaction occured when Carlos stirred and heated butanic acid and butanol. Chemical reaction – is when two or more substances interact to make a new substance. Before the reaction the mass of butanic acid was 9.78 g and the butanol was 8.22. After the reaction the mass of the butanic acid was 1.74 g and the butanol was 2.00 g Therefore a chemical reactions did occur.

This student used the data that the mass changed to determine that a chemical reaction occurred. In the substance and property explanation, students were less likely to use the inappropriate evidence to make the correct claim. In this question, liquid 1 and liquid 2 have the same mass, while liquid 1 and liquid 4 have the same density, color, and melting point. Consequently, a student who focuses on mass is less likely to make an accurate claim. For example, one student responded:

> Student E:
> No. None of the liquid was the same but liquid 1 and 4 would have been the same substance if their mass was the same.

By including mass, this student decided that liquids 1 and 4 were not the same substance. The use of inappropriate evidence results in an incorrect claim.

This provides an important lesson for both the design and evaluation of assessment items. It is important to consider the different information that can be used to construct the correct answer for a question. Although we consciously included inappropriate evidence in the assessment items because we were interested in how students would use the evidence, we did not consider that the inappropriate evidence could be used to construct the correct claim for the chemical reaction item. Since the inappropriate evidence was included in an open-ended item, students' written responses offered some insight into their use of the inappropriate evidence.

Designers need to be particularly careful when including inappropriate evidence in a multiple-choice item to consider how a student might use the evidence.

Differences in what counts as evidence. The question still remains why students were more likely to include inappropriate evidence in their explanations for chemical reactions (see Table 1). One possibility, is again, the wording of the question. For the chemical reaction explanation, students might have known that they were looking for whether "a change" occurred. Since all five measurements (density, melting point, mass, volume, and solubility) changed, students could use all of the measurements to make the claim so they might not have considered the appropriateness of each data point. For the substance and property explanation, students might have known that they were looking for "similar" measurements. In the assessment item, different measurements are the same for different pairs of liquids. Mass and color are the same for liquids 1 and 2, while density, color and melting point are the same for liquids 1 and 4. This difference may have encouraged students to think more deeply about the appropriateness of each data point for determining whether two liquids are the same substance.

Another possibility is that students' understanding of what counts as evidence for a chemical reaction was not as stronger as their understanding of what counts as evidence for two substances to be the same. Understanding chemical reactions builds off of their understanding of substance and properties and it may be more difficult for students. Students may have understood that mass and/or volume are not properties to differentiate substances yet they still thought they were signs of a chemical reaction. For example, below are one student's responses to both the substance and property question and the chemical reaction question.

> Student G:
> Substance and Property:
> Liquid 1 and 4 are of the same substance. Because all of the properties are the same, color, density, and melting point are the same. Mass is the same, but it does not count.
>
> Chemical Reaction:
> When the density and the mass changed and went up higher that showed that a chemical reaction occurred.

It is interesting that this student did not think that mass "counted" for determining whether two substances were the same, yet he thought it was important for determining whether a chemical reaction occurred. There were in fact 141 students who thought mass and/or volume were important for determining whether a chemical reaction occurred, but did not think that mass was important for determining whether two liquids are the same substance. Perhaps students thought that while mass and volume are not properties, they are still some how a sign of a chemical reaction. The majority of students, who used mass in their chemical reaction explanations, but not in their substance explanations, did not explicitly articulate why they were making that distinction. However, there were a couple of students who described why they thought it was important to include mass and volume as evidence for the chemical reaction explanation, but not for the substance and property explanation. The response below offers one example.

> Student H:
> Substance and Property:
> Liquid 1 and Liquid 4 are the same substances. They have the same density, 0.93 g/cm^3. They are both colorless. They both have the same melting point of -98° C. The only different about them is their mass, but mass is not a property because it varies with sample sizes. The evidence shows that Liquid 1 and Liquid 4 are the same substances because they have the same properties.
>
> Chemical Reaction:
> A chemical reaction did occur. Evidence of this is that neither of the beginning substances share the same amount of density with either of the end substances. Also, the melting points changed from –7.9 ° C and 89.5 ° C to –91.5 ° C and 0.0 ° C. Another piece of evidence is that the mass changed from 10.18 cm^3 and

> 10.15 cm^3 to 2.00 cm^3 and 2.00 cm^3. The solubility also changed. Because the mass and volume decreased so much, I think that gas formed. This data is evidence of a chemical reaction because properties changed.

This suggests that when the student read the data table for the chemical reaction explanation, she thought the differences in mass and volume told her about whether a gas formed. The student interpreted the data for after mixing as the total mass and volume of Layer A and Layer B instead of realizing that the text states that Carlos took a sample of Layer A and Layer B.

Students might have been confused by mass and volume because of the investigations they completed during the unit. Students performed a couple of investigations that included chemical reactions that produced a gas. For example, students combined sodium bicarbonate (baking soda) and calcium chloride (road salt) with a solution of phenol red in a sealed plastic bag and then observed three major changes: temperature change, color change, and the bag inflated with a gas (carbon dioxide)[5]. The transcript below is from a classroom discussion of this reaction. The transcript focuses on one group of three students who have just combined the substances in the plastic bag.

S1: Hey stop. Hey it turned yellow.
S2: It is changing colors.
S3: Mrs., Ms., Ms. Hill, it is changing to yellow now.
S1: It turned into yellow.
S2: Come on.
S3: Yeah. It is hot right here. Feel right there, Derek.
Teacher: Ok. What else is going on? We need to write down our observations. Yours is starting to get hot? Oh.
S2: There are bubbles. There is a temperature change.
Teacher: What's going on with the bag?
S2: It is shrinking. (pause). It is airing up. I mean.
Teacher: Write down our observations (Addressing the whole class).

---

[5] Although the students do not make this distinction, two different chemical reactions actually occur in this investigation: 1) sodium bicarbonate and calcium chloride form sodium chloride, calcium carbonate, carbon dioxide and water; 2) carbon dioxide, water and phenol red form hydrogen-carbonate ion and altered phenol red. Phenol red is an acid/base indicator and changes from red to yellow because of the altered acidity of the solution.

S3: It looks like -
S2: It is getting cold.
Teacher: Oh. So you are telling me that the color does not turn yellow? (Addressing a different group)
S2: And the bag is blown up.
Teacher: All right, wait we need to be writing this down. It started to bubble.
S2: There is fizz, temperature change (Quietly talking while writing).
Teacher: Wait a minute. Wait these bags are starting to get bigger to me.
S4: Yup. It is starting to be inflated.
Multiple Students: Laugh.
Teacher: Oh. I like that word.

Both the students and teacher discussed how the bag inflated or got bigger. Students associated this change in size with a chemical reaction. In retrospect, the curriculum did not clearly distinguish between the volume and mass of the chemical reaction as a whole system and the volume and mass of individual substances. Hence, students may have been confused by when mass and volume count as evidence. From their experiences with substances, students may have understood that mass and volume are not properties so they are inappropriate evidence to determine whether two liquids are the same substance. They may also have understood that mass is conserved in a closed system, but can change in an open system. Yet they may have been unclear of the role of mass and volume to determine whether a chemical reaction occurs. Students' responses for the chemical reaction explanation suggest that a number of students thought that a change in mass and volume counted as evidence for a chemical reaction.

Conclusion

Students do not typically construct strong explanations where they support their knowledge claims (Sadler, 2004). Yet constructing explanations can be a powerful way for students to actively construct knowledge. By engaging in an instructional unit where students received an explicit framework for scientific explanation, multiple opportunities to construct

explanations and support during those learning tasks, students created stronger explanations by the end of the unit. In post unit assessment tasks, students provided stronger claims and justification for those claims including evidence and reasoning. In contrast to our previous research (McNeill, et al., in review; McNeill, et al., 2003), we see that students reasoning scores started to approach their evidence scores by the end of the unit. These improved learning gains for reasoning may be the result of our revisions to the unit in which we made the reasoning component more explicit for students and provided more detailed scaffolds in the student investigation sheets.

Specifically in this study we examined when students used appropriate evidence and when they used inappropriate evidence. Similar to other research (Sandoval, 2003; Sandoval & Reiser, 1997), we found that students had difficulty including appropriate evidence to support their claims. At the end of the unit, a number of students still included inappropriate evidence in their explanations. We also found that students' ability to construct scientific explanations depended on the context. Students' ability to reason from data depends on the context, particularly in terms of students' prior understanding of the theoretical context (Masnick, Klahr, & Morris, this issue). Both students' understanding of the content knowledge and their understanding of the scientific inquiry practice can influence their ability to complete a practice. Individuals' lack of conceptual understanding can impede their ability to reason in science (Sadler, 2004). For the substance and property as well as the chemical reaction explanation, we found that students with stronger content understanding constructed stronger explanations and were less likely to use inappropriate evidence in their explanations. This suggests that strong content knowledge is important to appropriately take part in scientific inquiry practices such as accurately constructing scientific explanations. Students may be unable to apply their

understanding of a scientific inquiry practice to a context without an understanding of the particular science content. In our current research, we are exploring how both domain specific knowledge and general knowledge of scientific explanations influence students' ability to construct scientific explanations, as well as the roles of curriculum scaffolds and teacher practices in students' learning of both types of knowledge.

In this study, we found that students' use of evidence varied for the two different assessment tasks. Specifically, students were more likely to include inappropriate evidence in their explanation for the chemical reaction assessment task. We conjecture that there are two possible causes for students' use of inappropriate evidence for this task: the wording of the assessment task and difficulty knowing what counts as evidence for chemical reactions.

In assessing students' ability to construct explanations or analyze data, it is important to consider what knowledge is needed to accurately answer the assessment task. In constructing the chemical reaction explanation item, we did not consider that the inappropriate evidence could be used to support an accurate claim for the question. Project 2061 has created an analysis procedure for assessment items in which they determine the alignment to a learning goal based on whether the content is both necessary and sufficient to answer the question (Stern & Ahlgren, 2002). In the chemical reaction task, knowing that changes in mass and volume are not always evidence of a chemical reaction (e.g. could be the result of a phase change) was not necessary to construct the correct claim that a chemical reaction occurred. Rather students could believe that changes in mass and volume are evidence of a chemical reaction and actually construct the correct claim. This differed compared to the substance and property item where the mass data suggested that the wrong two liquids were the same substance, liquids 1 and 2. This suggests that

it is important to consider what knowledge is needed to construct the correct claim for an assessment task. Otherwise, an assessment item may not be testing the desired knowledge.

Students' use of inappropriate evidence in the chemical reaction item may also have been influenced by what they thought counted as evidence of a chemical reaction. When interpreting data, people take into consideration whether they can imagine a mechanism that might account for any patterns in the data (Koslowski, 1996). In the chemical reaction task, students may have attempted to come up with a mechanism for the decreasing mass and volume. In their previous experiments in class, they found that when chemical reactions produce gas that a change in mass and volume can occur. Connecting this classroom experience to the change in mass and volume in the chemical reaction assessment task may be why more students used inappropriate data for this task compared to the substance and property task.

Students associated "change" with chemical reactions and they could imagine a plausible mechanism to account for this change. It may be that students had a beginning understanding of chemical reactions that involved change, but had not yet differentiated what does and what does not change in a chemical reaction. Furthermore, understanding chemical reactions builds from an understanding of substance and properties. The chemical reaction assessment task requires more sophisticated thinking and links to other related knowledge structures, because it requires that students first have an understanding of properties and substance. This may make it more difficult for students to understand what counts as evidence for a chemical reaction to occur and may be why more students included inappropriate evidence in their chemical reaction explanation.

By having students construct explanations where they provide not only their claims, but also their evidence and reasoning, we obtained greater insight into student thinking. If this assessment task had been a multiple-choice item or only asked students to state whether a

chemical reaction occurred, we could not tell that a number of students were in fact using inappropriate evidence to create the claim. Having students construct scientific explanations can be an important tool to help make students thinking visible for both researchers and teachers. Encouraging students to articulate their evidence and reasoning provides researchers more information about how to revise instructional materials and provides teachers with important formative assessment. Formative assessments allow teachers to use the evidence from the assessment to change their instructional plans to better meet the needs of their students (Black, 2003).

Students' success in using evidence depends on both the content and context of the learning task. In previous iterations of revising the curriculum, we added activities and phenomena to specifically address that mass and volume are not properties and cannot be used to differentiate substances. We have not explicitly addressed why you would not rely on mass and volume to determine whether a chemical reaction occurred. In future revisions of the curriculum, we plan to address this student difficulty. We hope to include greater support during the unit to help students understand what counts as evidence for chemical reactions. Furthermore, we plan to revise the chemical reaction assessment task so that mass and volume can no longer be used to construct the correct claim. Although we plan to continue including inappropriate evidence in our items, we need to think more carefully about how students may use that evidence in their responses and what it means when they include inappropriate evidence**.**

We also need to continue providing students with practice to both use evidence in their explanations and critique other people's use of evidence in explanation. If our goal is to help students develop competent participation in a science infused world (McGinn & Roth, 1999), success on one learning or assessment task is not sufficient. Analyzing data and using data to

support claims is a complex task that varies depending on the context. Students need considerable practice to understand what counts as evidence to support knowledge claims and how that evidence changes depending on the content and context of the task.

Acknowledgements

The research reported here was supported in part by the National Science Foundation (REC 0101780 and 0227557). Any opinions expressed in this work are those of the authors and do not necessarily represent either those of the funding agency or the University of Michigan.

References

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

Bell, P., & Linn, M. (2000). Scientific arguments as learning artifacts: Designing for learning from the web with KIE. *International Journal of Science Education. 22 (8),* 797-817.

Black, P. (2003). The importance of everyday assessment. In J. M. Atkin & J. E. Coffey (Eds.), *Everyday assessment in the science classroom* (pp. 1-11). Arlington, VA: NSTA Press.

Blumenfeld, P. C., Marx, R. W., Patrick, H. Krajcik, J., & Soloway, E. (1997). Teaching for understanding. In B. J. Biddle, T. L. Good, & I. F. Goodson (Eds.), *International handbook of teachers and teaching* (pp. 819-878). Dordrecht, The Netherlands: Kluwer.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (Eds.). (2000). *How people learn: Brain, mind, experience, and school.* Washington, DC: National Research Council.

Brown, A. L., Ash, D., Rutherford, M., Nakagawa, K., Gordon, A., & Campione, J. C. (1993). Distributed expertise in the classroom. In G. Salomon (Ed.), *Distributed*

cognitions: Psychological and educational considerations (pp. 188-228). Cambridge: Cambridge University Press.

Chinn, C. A. & Brewer, W. F. (2001). Models of data: A theory of how people evaluate data. *Cognition and Instruction*, *19*, 323-393.

Crawford, B. A. (2000). Embracing the essence of inquiry: New roles for science teachers. *Journal of Research in Science Teaching, 37*(9), 916-937.

Driver, R., Asoko, H., Leach, J., Mortimer, E., & Scott, P. (1994). Constructing scientific knowledge in the classroom. *Educational Researcher, 23*(7), 5-12.

Driver, R., Newton, P. & Osborne, J. (2000). Establishing the norms of scientific argumentation in classrooms. *Science Education. 84 (3),* 287-312.

Harris, C. J., McNeill, K. L., Lizotte, D. L., Marx, R. W. & Krajcik, J. (in press). Usable assessments for teaching science content and inquiry standards. *Peers Matter*, 1 (1*)*.

Hogan, K. & Maglienti, M. (2001). Comparing the epistemological underpinnings of students and scientists' reasoning about conclusions. *Journal of Research in Science Teaching*. *38*(6). 663-687.

Jiménez-Aleixandre, M. P., Rodríguez, A. B., & Duschl, R. A. (2000). "Doing the lesson" or "doing science": argument in high school genetics. *Science Education, 84*, 757-792.

Kanari, Z. & Millar, R. (2004). Reasoning from data: How students collect and interpret data in science investigations. *Journal of Research in Science Teaching*. *31*(7). 748-769.

Koslowski, B. (1996). *Theory and evidence: The development of scientific reasoning.* Cambridge, MA: MIT Press.

Krajcik, J., Blumenfeld, P., Marx, R., & Soloway, E. (2000). Instructional, Curricular, and

Technological Supports for Inquiry in Science Classrooms. In J. Minstrell & E. v. Zee (Eds.), *Inquiring into Inquiry Learning and Teaching in Science* (pp. 283-315). Washington D.C.: AAAS.

Kuhn, D. (1993) Science as argument: Implications for teaching and learning scientific thinking. *Science Education, 77,* 319-338.

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

Lee, H.-S. & Songer, N. B. (2004, April). *Longitudinal knowledge development: Scaffolds for Inquiry.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Lehrer, R., & Schauble, L. (Eds.). (2002). *Investigating real data in the classroom: Expanding children's understanding of math and science*. New York: Teachers College Press.

Lemke, J. (1990). *Talking science: Language, learning, and values*. Norwood, NJ: Ablex Publishing Corporation.

Lizotte, D. J., McNeill, K. L., & Krajcik, J. (2004). Teacher practices that support students' construction of scientific explanations in middle school classrooms. In Y. Kafai, W. Sandoval, N. Enyedy, A. Nixon & F. Herrera (eds.), *Proceedings of the sixth international conference of the learning sciences* (pp. 310-317). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Masnick, A. M., Klahr, D. & Morris, B. J. (this volume). Separating signal from noise: Children's understanding of error and variability in experimental outcomes. In Lovett, M & Shah, P (Eds.) *Thinking with Data: the Proceedings of the 33rd Carnegie Symposium on Cognition.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

McGinn, M. K. & Roth, W-M. (1999). Preparing students for competent scientific practice: Implications of recent research in science and technology studies. *Educational Researcher, 28*(3) 14-24.

McNeill, K. L., Harris, C. J., Heitzman, M., Lizotte, D. J., Sutherland, L. M., & Krajcik, J. (2004). How can I make new stuff from old stuff? In J. Krajcik & B. J. Reiser (Eds.), *IQWST: Investigating and questioning our world through science and technology*. Ann Arbor, MI: University of Michigan.

McNeill, K. L., Lizotte, D. J, Harris, C. J., Scott, L. A., Krajcik, J., & Marx, R. W. (2003, March). *Using backward design to create standards-based middle-school inquiry-oriented chemistry curriculum and assessment materials*. Paper presented at the annual meeting of the National Association for Research in Science Teaching, Philadelphia, PA.

McNeill, K. L., Lizotte, D. J, Krajcik, J., & Marx, R. W. (in review). Supporting students' construction of scientific explanations by fading scaffolds in instructional materials.

Metz, K. E. (2000). Young children's inquiry in biology: Building the knowledge bases to empower independent inquiry. In J. Minstrell & E. H. van Zee (eds.), *Inquiry into inquiry learning and teaching in science* (pp. 371-404). Washington, DC: American Association for the Advancement of Science.

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

Reiser, B. J., Krajcik, J., Moje, E. B., & Marx, R. W. (2003, March). *Design strategies for developing science instructional materials*. Paper presented at the Annual Meeting of the National Association for Research in Science Teaching, Philadelphia, PA.

Reiser, B., Tabak, I., Sandoval, W., Smith, B., Steinmuller, F., & Leone, A. (2001). BGuILE: Strategic and conceptual scaffolds for scientific inquiry in biology classrooms. In S.M. Carver & D. Klahr (Eds.), *Cognition and instruction: Twenty-five years of progress* (pp. 263-305). Mahwah, NJ: Erlbaum.

Rivet, A. E. & Krajcik, J. S. (2004). Achieving standards I n urban systemic reform: An example of a sixth grade project-based science curriculum. *Journal of Research in Science Teaching*. *41*(7). 669-692.

Sadler, T. D. (2004). Informal reasoning regarding socioscientific issues: A critical review of research. *Journal of Research in Science Teaching*. *41*(5). 513-536.

Sandoval, W. (2003). Conceptual and epistemic aspects of students' scientific explanations. *The Journal of the Learning Sciences, 12*(1), 5-51.

Sandoval, W. A. & Reiser, B. (1997, March). *Evolving explanations in high school biology*. Paper presented at the Annual Meeting of the American Educational Research Association. Chicago, IL.

Singer, J., Marx, R., Krajcik, J., & Chambers, J. (2000). Constructing Extended Inquiry Projects: Curriculum Materials for Science Education. *Educational Psychologist*. 35(3), 165-178.

Salomon, G., D. N. Perkins, & T. Globerson. (1991). Partners in cognition: Extending human intelligence with intelligent technologies. *Educational Researcher. 20*(2)*:* 2-9.

Stern, L. & Ahlgren, A. (2002). Analysis of students' assessments in middle school curriculum materials: Aiming precisely at benchmarks and standards. *Journal of Research in Science Teaching, 39*(9), 889-910.

Tabak, I., & Reiser, B. J. (1997). Complementary roles of software-based scaffolding and teacher-student interactions in inquiry learning. In R. Hall, N. Miyake & N. Enyedy

(Eds.), *Proceedings of Computer Support for Collaborative Learning '97* (pp. 289-298). Toronto, Canada.

Toulmin, S. (1958). *The uses of argument.* Cambridge, UK: Cambridge University Press.

White, B., & Frederiksen, J. (1998). Inquiry, modeling, and metacognition: Making science accessible to all students. *Cognition and Instruction, 16*(1), 3-118.

Wiggins, G., & McTighe, J. (1998). *Understanding by Design.* Alexandria, VA: Association for Supervision and Curriculum Development.

Zembal-Saul, C., Munford, D., Crawford, B., Friedrichsen, P. & Land, S. (2002). Scaffolding preservice science teachers' evidence-based arguments during an investigation of natural selection. *Research in Science Education*, *32 (4),* 437-465.

Appendix A: Sample Multiple-Choice Items

1. To determine if a chemical reaction occurred, you should measure and compare which of the following?
   A. volume of the materials
   B. shape of the products
   C. properties of the substances
   D. mass of the reactants

5. Which of the following is an example of a chemical reaction?
   A. mixing lemonade powder with water
   B. burning marshmallows over a fire
   C. melting butter in a pan
   D. boiling water on a stove

12. A property is
   A. determined by the amount of a substance.
   B. made of one type of substance.
   C. a process to make a new substance.
   D. a characteristic of a substance.

3. A student found 2 green powders that look the same. He wants to figure out if the 2 powders are the same or different substances. Which of the following is the best method to use?
   A. Measure the mass, volume, and temperature of each powder and compare.
   B. Combine both green powders and see if there is a chemical reaction.
   C. Mix the 2 green powders together and then test the properties.
   D. Determine the density, solubility, and melting point of each powder and compare.

Appendix B: Scientific Explanation Items

Substance and Property Explanation:

Examine the following data table:

	Density	**Color**	**Mass**	**Melting Point**
Liquid 1	0.93 g/cm^3	no color	38 g	-98 °C
Liquid 2	0.79 g/cm^3	no color	38 g	26 °C
Liquid 3	13.6 g/cm^3	silver	21 g	-39 °C
Liquid 4	0.93 g/cm^3	no color	16 g	-98 °C

Write a **scientific explanation** that states whether any of the liquids are the same substance.

Chemical Reaction Explanation:

Carlos takes some measurements of two liquids — butanic acid and butanol. Then he stirs the two liquids together and heats them. After stirring and heating the liquids, they form two separate layers — layer A and layer B. Carlos uses an eyedropper to get a sample from each layer and takes some measurements of each sample. Here are his results:

		Measurements				
		Density	Melting Point	Mass	Volume	Solubility in water
Before stirring & heating	Butanic acid	0.96 g/cm^3	-7.9 ˚C	9.78 g	10.18 cm^3	Yes
	Butanol	0.81 g/cm^3	-89.5 ˚C	8.22 g	10.15 cm^3	Yes
After stirring & heating	Layer A	0.87 g/cm^3	-91.5 ˚C	1.74 g	2.00 cm^3	No
	Layer B	1.00 g/cm^3	0.0 ˚C	2.00 g	2.00 cm^3	Yes

Write a **scientific explanation** that states whether a chemical reaction occurred when Carlos stirred and heated butanic acid and butanol.

Appendix C: Specific Rubrics

Specific Rubric for Substance and Property Scientific Explanation

Component	Level		
Claim – A statement or conclusion that answers the original question/problem	0	1	2
	Does not make a claim, or makes an inaccurate claim. --- States none of the liquids are the same or specifies the wrong solids.	Makes an accurate but incomplete claim. --- Vague statement, like "some of the liquids are the same."	Makes an accurate and complete claim. --- Explicitly states "Liquids 1 and 4 are the same substance."
Evidence – Scientific data that supports the claim. The data needs to be appropriate and sufficient to support the claim.	0	1 & 2	3
	Does not provide evidence, or only provides inappropriate evidence (Evidence that does not support claim). --- Provides inappropriate data, like "the mass is the same" or provides vague evidence, like "the data table is my evidence."	Provides appropriate, but insufficient evidence to support claim. May include some inappropriate evidence. --- Provides 1 or 2 of the following pieces of evidence: the density, melting point, and color of liquids 1 and 4 are the same. May also include inappropriate evidence, like mass.	Provides appropriate and sufficient evidence to support claim. --- Provides all 3 of the following pieces of evidence: the density, melting point, and color of liquids 1 and 4 are the same.
Reasoning – A justification that links the claim and evidence and includes appropriate and sufficient scientific principles to defend the claim and evidence.	0	1, 2 & 3	4
	Does not provide reasoning, or only provides reasoning that does not link evidence to claim. --- Provides an inappropriate reasoning statement like "they are like the fat and soap we used in class" or does not provide any reasoning.	Repeats evidence and links it to the claim. May include some scientific principles, but not sufficient. --- Repeats the density, melting point, and color are the same and states that this shows they are the same substance. Or provides an incomplete generalization about properties, like "mass is not a property so it does not count."	Provides accurate and complete reasoning that links evidence to claim. Includes appropriate and sufficient scientific principles. --- Includes a complete generalization that density, melting point, and color are all properties. Same substances have the same properties. Since liquids 1 and 4 have the same properties there are the same substances.

Appendix C: Specific Rubrics

Specific Rubric for Chemical Reaction Scientific Explanation

Component	Level		
Claim – A statement or conclusion that answers the original question/problem	0 Does not make a claim, or makes an inaccurate claim. --- States that a chemical reaction did not occur.	Does not apply to this learning task.	1 Makes an accurate and complete claim. --- States that a chemical reaction did occur.
Evidence – Scientific data that supports the claim. The data needs to be appropriate and sufficient to support the claim.	0 Does not provide evidence, or only provides inappropriate evidence (Evidence that does not support claim). --- Provides inappropriate data, like "the mass and volume changed" or provides vague evidence, like "the data shows me it is true."	1 & 2 Provides appropriate, but insufficient evidence to support claim. May include some inappropriate evidence. --- Provides 1 or 2 of the following pieces of evidence: Butanic acid and butanol have different solubilities, melting points, and densities compared to Layer A and Layer B. May also include inappropriate evidence, like mass or volume.	3 Provides appropriate and sufficient evidence to support claim. --- Provides all 3 of the following pieces of evidence: Butanic acid and butanol have different solubilities, melting points, and densities compared to Layer A and Layer B. May also include inappropriate evidence, like mass.
Reasoning – A justification that links the claim and evidence and includes appropriate and sufficient scientific principles to defend the claim and evidence.	0 Does not provide reasoning, or only provides reasoning that does not link evidence to claim. --- Provides an inappropriate reasoning statement like "a chemical reaction did not occur because Layers A and B are not substances" or does not provide any reasoning.	1, 2, 3 & 4 Repeats evidence and links it to the claim. May include some scientific principles, but not sufficient. --- Repeats the solubility, melting point, and density changed, which show a reaction occurred. Or provides either A or B: A. A chemical reaction creates new or different substances OR B. Different substances have different properties.	5 Provides accurate and complete reasoning that links evidence to claim. Includes appropriate and sufficient scientific principles. --- Includes a complete generalization that: A. A chemical reaction creates new or different substances AND B. Different substances have different properties.

Table 1: Developing Learning Performances

Content Standard	Scientific Practice Standard	Learning performance
When substances interact to form new substances, the elements composing them combine in new ways. In such recombinations, the properties of the new combinations may be very different from those of the old (AAAS, 1990, p.47).	Develop…explanations… using evidence. (NRC, 1996, A: 1/4, 5-8) Think critically and logically to make the relationships between evidence and explanation. (NRC, 1996, A: 1/5, 5-8)	Students construct scientific explanations stating a claim whether a chemical reaction occurred, evidence in the form of properties, and reasoning that a chemical reaction is a process in which old substances interact to form new substances with different properties than the old substances.

Table 2: Participants from the 2003-2004 School Year

2003-2004 School Year			
Site	Urban	Suburban	Total
Schools	7	1	8
Teachers	7	3	10
Classrooms	29	5	34
Students	955	79	1034

Figure 1: Overall Student Learning Gains (n=700)

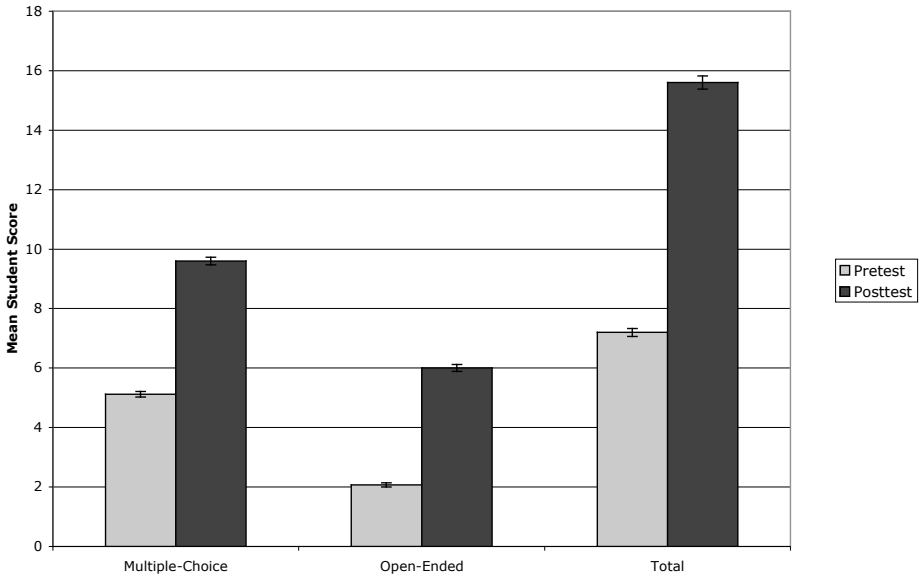

Figure 2: Student Learning Gains for Explanations (n=700)

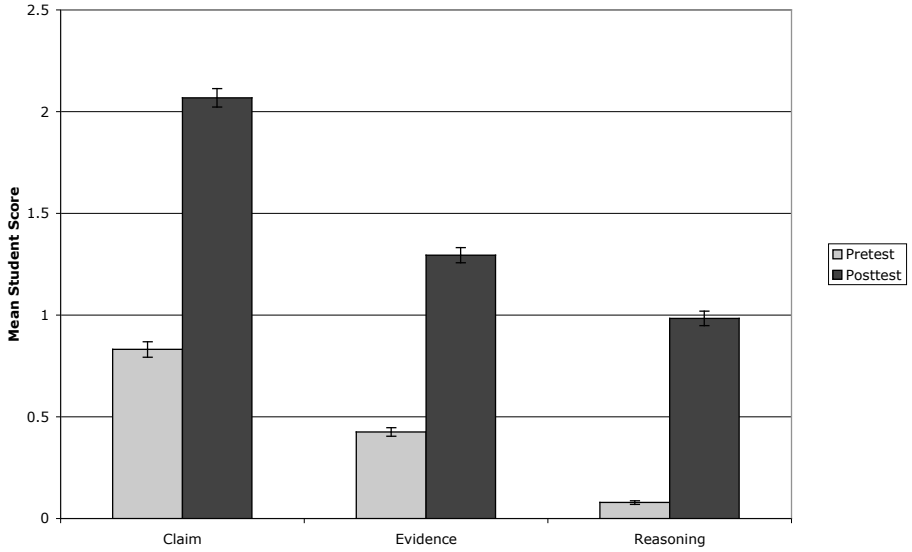

Figure 3: Learning Gains for Substance & Property Explanations (n= 700)

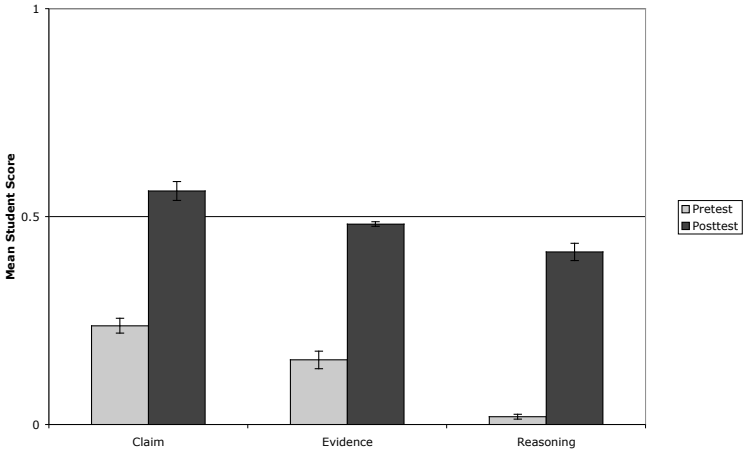

Figure 4: Learning Gains for Chemical Reaction Explanation (n=700)

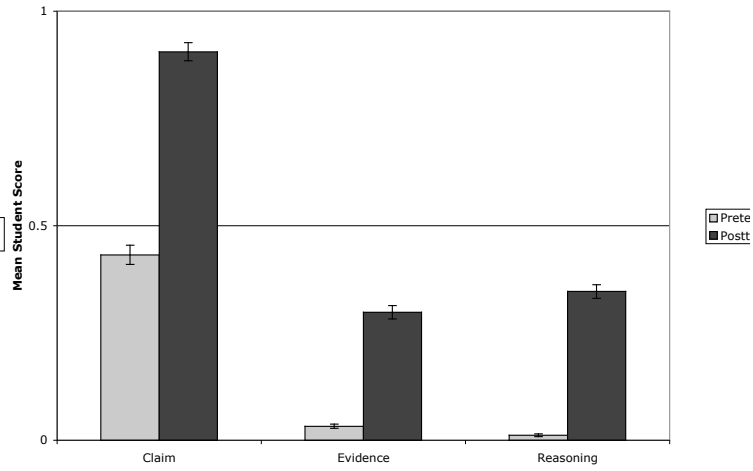

Table 3: Students' Use of Inappropriate Evidence in Explanations (n = 700)

			Substance and Property Explanation		Total
			0[a]	1[b]	
Chemical Reaction Explanation	0[a]	Count	436	80	516
		% of Total	62.3%	11.4%	73.7%
	1[b]	Count	141	43	184
		% of Total	19.9%	6.4%	26.3%
Total		Count	579	124	703
		% of Total	82.1%	17.9%	100.0%

a. 0 = Student did not use inappropriate evidence
b. 1 = Student did use inappropriate evidence

Figure 5: Substance and Property Claim and Reasoning Scores By Students' Use of Evidence

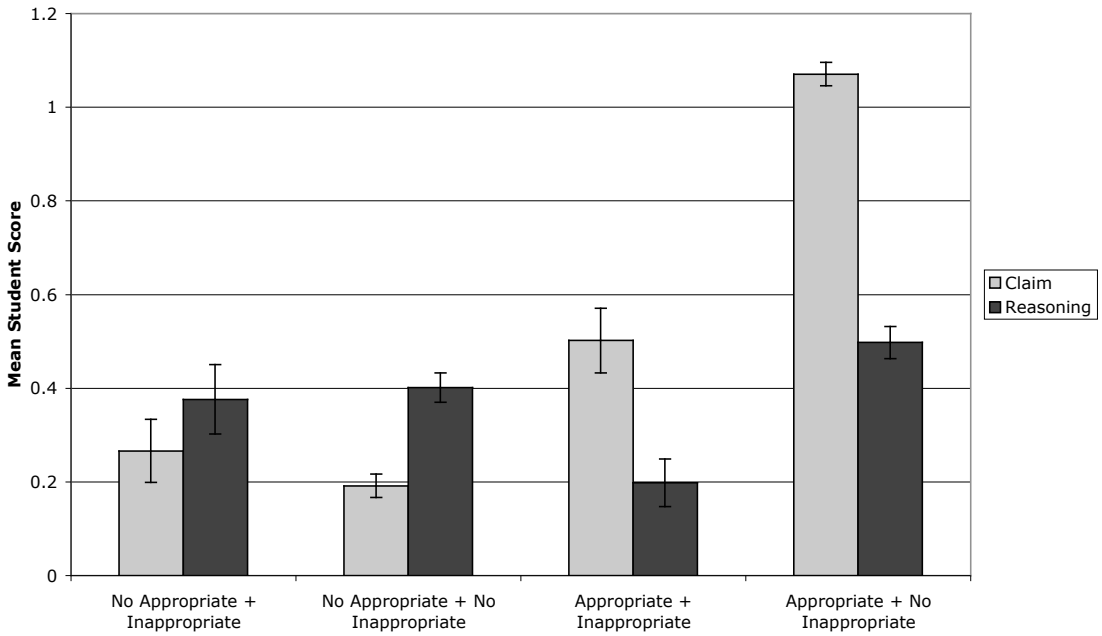

Figure 6: Substance and Property Multiple-Choice Scores By Students' Use of Evidence

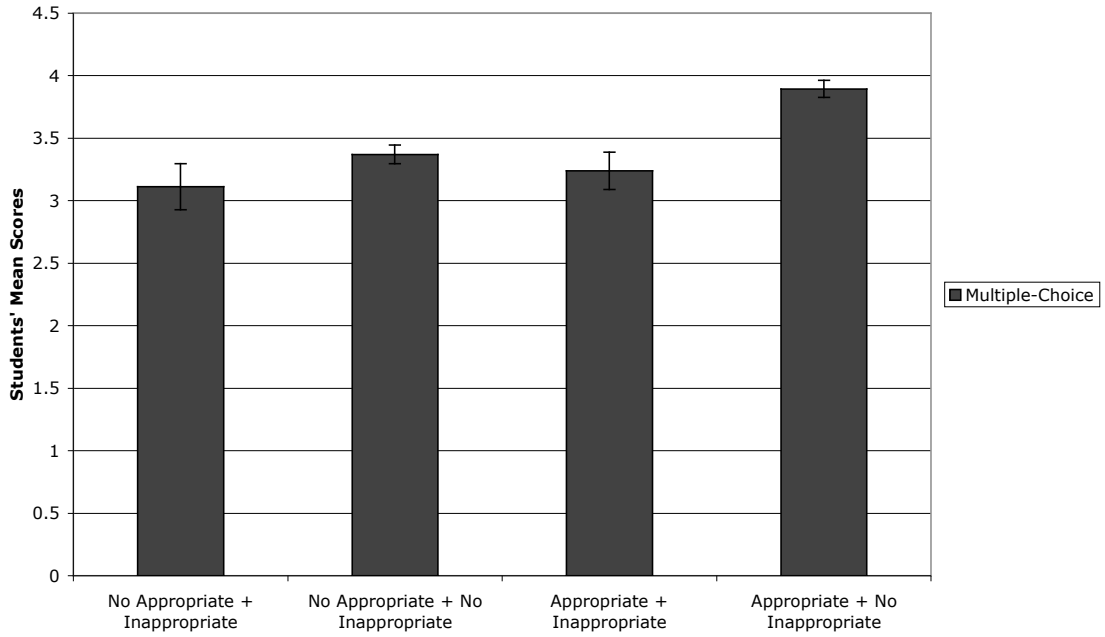

Figure 7: Chemical Reaction Claim and Reasoning Scores By Students' Use of Evidence

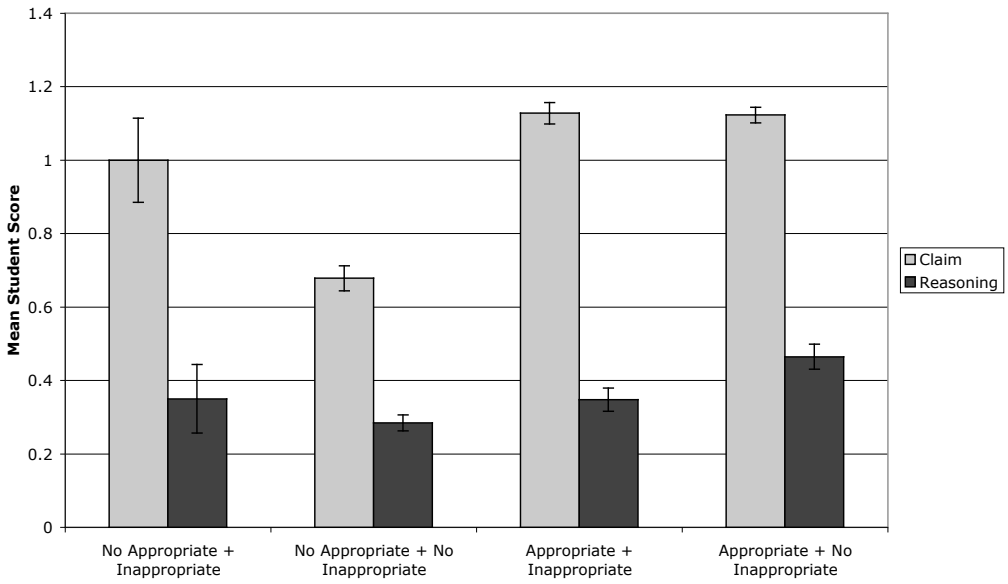

Figure 8: Chemical Reaction Multiple-Choice Scores By Students' Use of Evidence