Winter 1-17-2014

"What Does This Graph Mean?" Formative Assessment With Science Inquiry to Improve Data Analysis

Andrea Dawn Leech
Portland State University

“What Does This Graph Mean?”

Formative Assessment With Science Inquiry to Improve Data Analysis.

by

Andrea Dawn Leech

A thesis submitted in partial fulfillment of the
requirements for the degree of

Master of Science in Teaching
in
General Science

Thesis Committee:
William Becker, Chair
Emily Saxton
Nicole Rigelman

Portland State University
2013

Abstract

This study investigated the use of formative assessment to improve three specific data analysis skills within the context of a high school chemistry class: graph interpretation, pattern recognition, and making conclusions based on data. Students need to be able to collect data, analyze that data, and produce accurate scientific explanations (NRC, 2011) if they want to be ready for college and careers after high school. This mixed methods study, performed in a high school chemistry classroom, investigated the impact of the formative assessment process on data analysis skills that require higher order thinking. We hypothesized that the use of evaluative feedback within the formative assessment process would improve specific data analysis skills. The evaluative feedback was given to the one group and withheld from the other for the first part of the study. The treatment group had statistically better data analysis skills after evaluative feedback over the control. While these results are promising, they must be considered preliminary due to a number of limitations involved in this study.

Key words: formative evaluation, formative assessment, science inquiry, chemistry, high school, and science education

Dedication

To my husband Chris and my son Tyler for all their support during this process

Table of Contents

List of Tables

List of Figures

Introduction

The National Research Council (2011) recently published a book titled, *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas* (*A framework*). *A framework* states that understanding science is a basic requirement for navigating life in the current technological age. *A framework* identifies three dimensions of science curriculum that are required for science understanding. One of the three dimensions is a set of eight scientific and engineering practices. Two of these eight scientific and engineering practices are related to data analysis. These two practices are "analyzing and interpreting data" (p. 61), and "engaging in argument from evidence" (p. 71). *A framework* defines "analyzing and interpreting data" as including the following set of data analysis skills: recognizing patterns, evaluating conclusions, exploring relationships with graphs, and inferring relationships. The book defines "engaging argument from evidence" to include the following data analysis skills: supporting a claim with data and discussing that claim with evidence and reasoning.

Multiple studies indicate that students struggle with data analysis skills. Hug and McNeill (2008) found issues with recognizing patterns, inferring relationships and constructing scientific arguments, which is also referred to as scientific explanation. Sandoval (2003) and Driver, Newton and Osborne (2000) found problems with scientific explanations. Shah and Hoeffner (2002) found problems with exploring relationships using graphs. Kanari and Millar (2004) outline issues with recognizing patterns in the data. In essence, students struggle with every data analysis skill that is mentioned by the National Research Council.

Because students struggle with data analysis skills, teachers must find new ways to address these topics if they want to ensure students master these skills. Vygotsky (2011), states that students learn best when they operate inside their zone of proximal development (ZPD). The ZPD consists of challenges that are beyond a student's ability to complete alone, but can be completed with help. To maximize learning in this zone, a student is paired with a mentor or instructor who provides a small amount of assistance to the student. The assistance provided by the mentor expands the student's understanding. Students then learn instead of struggling.

Vygotsky does not define exactly what a mentor's assistance has to look like. He only states that mentors provide assistance to facilitate learning. More recent work (Van Der Stuyf, 2002) has named this type of assistance scaffolding. Various forms of scaffolding could be used such as modeling performance, clearly defining expectations, providing direction on reaching goals and findings ways to motivate students.

Black and Wiliam (1998) outline one highly effective form of assistance, formative assessment (FA). FA begins by assessing student understanding with respect to learning targets. This would, in theory, help a teacher identify each student's particular ZPD. The teacher acts as a mentor by providing small amounts of targeted assistance in the form of evaluative feedback. This feedback helps students understand expectations and gives them direction on reaching learning targets. After students obtain feedback, they are given opportunities to work within their ZPD. The work that the students are given consists of improving their own work to reach the learning goals. Multiple studies have shown that this process

improves learning in a number of different contexts (Black & Wiliam, 1998; Torrance, 2007; Shute, 2008; Shavelson et al. 2008; Hume & Coll, 2009), but the studies on FA learning gains have focused on mastery of content knowledge only. Studies have not been performed that apply formative assessment to the two key data analysis practices that were listed above.

This raises the question, "How do we use FA to improve data analysis skills?" As was stated above, the first step is determining student understanding with some form of assessment. Not all assessments are effective at evaluating all cognitive processes (Ruiz-Primo & Shavelson, 1996). Therefore, it is important to ensure an appropriate assessment by determining what kind of cognitive processes are being targeted. According to Leighton (2011), the data analysis skills described by the National Research Council in *A framework* (2011) are higher order thinking. For example, Leighton specifies that the following activities use higher order thinking skills: inquiry investigations, using appropriate techniques to analyze and interpret data, developing inferences, predictions and arguments aligned with evidence, and substantiating/evaluating claims. Thus, formative assessment of data analysis skills must begin with an assessment of higher order thinking.

Higher order thinking is difficult to assess (Yeh, 2001; Ruiz-Primo & Shavelson, 1996). It needs to be done in a way that makes student thinking explicit (Furtak, & Ruiz‐Primo, 2008) through performance items, short-answer completion questions, and projects (Leighton, 2011). One way to make student thinking explicit is to have students write a work sample or lab report in

conjunction with a science inquiry project (Ruiz-Primo, Li, Tsai, & Schneider, 2010). This provides a window into student thinking and provides a number of additional benefits to the students. For example, science inquiry has been shown to improve content understanding across multiple schools and teachers (Marx et al., 2004), higher order thinking skills (Hofstein, Shore & Kipnis, 2004; Berg, Bergendahl, Lundberg, & Tibell 2003), and attitudes towards science (Hofstein et al., 2004).

In summary, the National Research Council indicates that data analysis skills are important, but students struggle with these skills. FA improves learning, but has predominantly been applied to content knowledge instead of data analysis skills. This creates a gap in the literature. Data analysis skills utilize higher order thinking, which can be assessed using short-response questions in an assessment or as part of the lab report write up at the culmination of a science inquiry project. To close the gap in the literature, a study was designed to investigate the impact of formative assessment on data analysis skills using both the lab report that was created at the culmination of a science inquiry project and a set of formative assessment instruments that consisted of short answer questions.

A set of formative assessments instruments were created for this study and vetted by an expert panel. The instruments consisted of short response questions based on a short science scenario that included a graph. The instruments were used to assess data analysis skills at different points in this study: before either class received treatment, after one class received treatment and after both classes received treatment. After the initial pre-assessment, students were taught data analysis skills using a variety of instructional strategies. This instruction included

modeling, science inquiry, and interactive discussion. Students also received evaluative feedback associated with formative assessment and were given time to incorporate that feedback into their work. Afterwards, student's data analysis skills were assessed using a FA instrument that was parallel to the pre-assessment.

This study was done on classes of high school chemistry students from an ethnically diverse school in the Portland Metro area. Class A received the treatment while class B is used as a control. Halfway through the experiment, the classes were switched. Class A acted as the control while class B received the treatment. Learning gains were determined using a series of three assessments that were created by the researcher and reviewed for content validity by a panel of 8 experts. Afterwards, a series of cognitive interviews was done using a "Think-aloud" assessment to better understand student thinking.

Literature Review

Teaching data analysis skills

It is well documented in the literature that students have problems with data analysis. Hug and McNeill (2008) looked at 27 studies across different grade levels and contexts and identified multiple issues with data analysis: data measurement, limitations of data, data source, data manipulation, patterns/inferences, conclusions, consideration of content knowledge, and use of everyday examples or analogies. Shah and Hoeffner (2002) reviewed 83 articles and found that students had difficulty interpreting graphs and understanding patterns in those graphs. Kanari and Millar (2004) looked at the ways that 60 middle grade students, ages 10-14, collected and interpreted data. They found that students assumed relationships that did not exist within graphs and ignored data that contradicted their preconceived notions about causation.

Schauble, Glaser, Duschl, Schulze and John (1995) followed videotaped instructional activities in five sixth grade science classrooms and interviewed 21 of the students involved. The researchers found only 14% of the students understood the variables in the experiment. Only 29% of the students understood the primary relationship shown by the numbers on the chart and only 14% conceptually understood what they were doing. The researchers concluded that students did not adequately recognize the patterns in the data or draw appropriate conclusions. These studies show us that students continue to have problems with data analysis skills in spite of instruction. We must examine the methods of instruction to understand how teaching efforts fail.

One of the reasons students have problems with data analysis is that they are expected to transfer knowledge from other subjects like math. Keiler (2007) interviewed 60 students aged 15-16 and 11 teachers to understand where students learned data analysis skills. Students felt like analysis and interpretation skills (like data measurement and manipulation) were learned in math while planning and evaluation skills (like scientific explanation and patterns) were learned in science. Teachers recognized that students who struggled with analysis and interpretation skills had weak math skills. The teachers felt that students who struggled the most with data analysis were not transferring their math skills to science.

Why is this transfer so hard? Many data analysis skills require higher order thinking.

For example, *A framework* (NRC, 2011) lists the following as targeted data analysis skills: recognizing patterns, evaluating conclusions, exploring relationships with graphs, inferring relationships, supporting a claim with data and discussing that claim with evidence and reasoning. In the revised Bloom's taxonomy, Anderson, Krathwohl, and Bloom (2001) indicate that a majority of these skills involve higher order thinking. Leighton (2011) indicates that higher order thinking must be assessed using short response questions or projects. Furtak and Ruiz-Primo (2008) go further and state that student thinking must be explicit in order for teachers to adequately assess higher order thinking.

Unfortunately, many chemistry textbooks do not offer the kinds of questions that are needed to assess higher order thinking. Overman, Vermunt, Meijer, Bulte, and Brekelmans (2012) reviewed 971 chemistry questions from four different

curriculums. While it could be construed that data analysis skills were explored in these curriculum, the majority of the questions were applying formulas to gain mathematical answers instead of the higher order analysis described in *A Framework* (2011). Similar results were found by Davila and Talanquer (2009) when they looked at college textbooks. The majority of the questions fell into Blooms Taxonomy under Application or Analysis. The few analysis questions they found required students to make inferences or predictions, which is one of the highlighted data analysis skills. Unfortunately, the other data analysis skills outlined in *A Framework* (NRC, 2011) were not addressed.

If textbooks do not ask questions that force students to use higher order thinking, students will take shortcuts that allow them to avoid using higher order thinking. Lovett and Chang (2007) studied ten undergraduate students that were taught an explicit data analysis methodology in a college statistics class. These students rejected the systematic methodology they were taught in favor of a "guess and test" method. In many cases student work was missing appropriate evaluations about the strength of their conclusions and evidence. Evaluating the strength of conclusions is listed as one of the important practices that the NRC associated with data analysis (2011). In a follow-up study, Lovett and Chang (2007) looked at 52 participants who had completed their bachelor's degree to study the impact of practice problem wording on student learning. Students preferentially used the written cues in the problem over applying the crucial data analysis skills that the problem was designed to help them practice. A similar finding was reported by McNeill and Krajcik (2007). Their study involved 1034 students across eight school

districts and investigated issues with scientific explanations. When students got the correct answer using faulty logic, the students did not gain a deep level of understanding. Dunbar, Fugelsang and Stein (2007) indicate that getting the correct answer through faulty logic may in fact strengthen misconceptions, which interfere with deeper understanding. In essence, questions need to be carefully worded to ensure that students engage critical thinking skills instead of using queues and hints from the problem that allow them to short-cut the thinking process. Since most textbook questions are not worded to force students to use higher order thinking, teachers must move beyond textbook questions to evaluate and engage higher order thinking skills.

In summary, students struggle with data analysis skills (Hug & McNeill, 2008; Shah & Hoeffner, 2002; Kanari & Millar, 2004; Schauble et al., 1995). There are several reasons for this. First, students struggle with transferring these skills from math to science (Keiler, 2007). Second, data analysis skills often involve higher order thinking (NRC, 2011; Leighton, 2011). Assessment of higher order thinking skills should be done with short answer questions or inquiry type projects (Ruiz-Primo & Shavelson, 1996; Furtak & Ruiz-Primo, 2008; Ruiz-Primo & Furtak, 2007; Leighton, 2011). A majority of textbook questions do not require higher order thinking skills (Overman et al., 2012; Davila & Talanquer, 2009). If students are not being assessed on higher order thinking, they will take shortcuts that allow them to avoid practicing higher order thinking (Lovett and Chang, 2007; McNeill & Krajcik, 2007). Teachers must shift their practice from using end of the chapter textbook

questions to more short answer questions that expose thinking and inquiry type projects.

Using science inquiry

One method of instruction that is currently being used in the chemistry classroom is the lab, but not all lab experiences are the same. In 1982, Hofstein and Lunetta concluded that the methods being used in laboratory experiences were crucial to obtaining positive learning outcomes. Hofstein's follow-up work in 2004 indicated that science inquiry produced more indications of higher order thinking than traditional expository experiences, where everything was scripted. Not only did this improve thinking skills and understanding, but it also improved student self-efficacy and positive attitudes towards science. Similar results were found with middle school students (Marx et al., 2004) and college students (Berg et al., 2003). This indicates that the way to improve higher order data analysis skills like explanation is to use inquiry.

What is science inquiry? Bell, Smetana and Binns (2005) define science inquiry as "an active learning process in which students answer research questions through data analysis" (p. 30). Wheeler and Bell (2012) split this definition into three different kinds of inquiry: structured, guided and open-ended. In structured inquiry, the question and procedures are known, but the solution is unknown. Guided inquiry provides the question, but requires the student to come up with the procedures and the solution. Open-ended inquiry means that the student supplies the question, the procedure and the question.

Even when these guidelines are followed, the result may not be inquiry. Teachers must instruct in an open manner that facilitates inquiry instruction (Hofstein & Lunetta, 1982). Bell et al. (2005) indicate that instruction needs to start with a question that must be answered by student observations and student analysis of data rather than a summary of findings from other scientists. Findings are can be written up in a lab report, discussed as a class or presented to a class by a small group. This process makes student thinking explicit, which is key to assessing higher order thinking (Furtak & Ruiz-Primo, 2008).

Science inquiry is a skill that must be learned (Bell et al., 2005; Berg et al., 2003; Wheeler & Bell, 2012). Wheeler and Bell (2012) suggest scaffolding inquiry by starting students with confirmatory or structured inquiry and slowly working towards open-ended inquiry. Sandoval (2003) also looked at scaffolding inquiry instruction. Sandoval looked at writing samples of 69 students in three high school biology classes studying evolution for four weeks. Students were asked to construct scientific explanations based on several sources of data. They found that explicit instruction on how to include scientific data in scientific explanations improved both the coherence and quality of these explanations. Keys, Hand, Prain, and Collins (1999) found feedback useful in helping students learn inquiry. Keys et al. studied two 8th grade classes during an eight-week stream study. These students were given explicit instruction on what was required and the rubric on how they were to be graded. Their performance did not improve until students received feedback on their poor performance in conjunction with an explanation of a successful paper.

Science inquiry labs improve learning, and thinking skills (Hofstein, 2004) when compared to traditional expository labs. There are three kinds: structured, guided and open-ended (Wheeler & Bell, 2012). In all types of science inquiry, the lab begins with a question. Students must use their own observations and data analysis skills to answer the question (Bell et al., 2005). Because it is more involved than traditional expository labs, students need scaffolding to learn science inquiry (Sandoval, 2003; Wheeler & Bell, 2012) this can be done by working through the different stages of inquiry starting with structured inquiry and ending with open-ended inquiry. Learning science inquiry is aided by feedback and opportunities to improve work (Keys et al., 1999). This is similar to the evaluative feedback and description of success that is part of formative assessment.

Formative assessment

One might ask, "What formative assessment is and how does it work?" The National Research Council (2000) report "How People Learn: Brain, Mind, Experience, and School," emphasized that the preponderance of the research finds that students have preconceived ideas about how the world works, and recommends that instructors relate new material to these initial preconceptions in order for students to shift their thinking and fully grasp new ideas. This applies not only to content material but also to data analysis skills and the use of data to make evidence-based claims. Part of the problem is that textbooks often outline rules for collecting, analyzing, and drawing valid conclusions from data, but do not tie these data analysis skills to common students preconceptions (NRC, 2000). If data

analysis is not tied to analyzing student preconceptions, students will not use data analysis to inform their thinking. The result is that students discount data that contradicts their personal preconceptions even though that data is valid (Kanari & Millar, 2004). Teachers need to bridge the gap between student understanding and application if teachers want students to progress beyond incorrect preconceptions. One way to do this is through formative assessments (FA).

Black and Wiliam (1998) reviewed 250 articles on FA and found significant improvements in content knowledge when four actions were performed: clearly stating the learning goals, identifying a learner's current level of knowledge, giving the learner feedback about gaps in knowledge and helping that student understand how to close the gap in their understanding. They cited papers where students improved math skills, obtained mastery over certain course objectives, and increased depth of knowledge. They concluded that implementing the four pieces of FA would help students refocus their efforts on learning and improve all aspects of learning.

Recently, FA gains have been questioned, but much of this stems from an inconsistent definition of FA. Dunn and Mulvenon (2009) attempted to do a statistical analysis of the FA literature. Instead, they found multiple definitions of FA. Bennett (2011) reported similar problems indicating confusion in the literature between FA as an instrument and as a process. When FA is viewed solely as an instrument, students don't always gain a boost in achievement. Torrance (2007) looked at 260 students who were involved in college or post high school vocational education. He found that when teachers focused on FA as an instrument instead of a

process, students did not achieve the learning gains expected. Hume and Coll (2009) followed up by working with 15-16 year olds studying science in New Zealand. They found that there were two components missing when educators viewed FA as a tool instead of a process: "feedback requiring students to respond … by [doing] further work," and self-assessment (p. 270). Shute (2008) found similar findings. She reviewed over 170 articles and found that feedback was a key part of formative assessment and that the complexity, timing and length all played roles in the success of these types on interventions. To be effective, FA must be a process and not just an instrument.

One key component of this process is evaluative feedback. Lee, Woods, and Tonissen (2011) looked at 70 students in a 3rd year undergraduate biological science laboratory. They introduced writing activities followed by formative feedback from teachers and peers and interviewed students to see the impact. Students used this feedback to improve subsequent writing exercises that were also reviewed by the teacher and peers. At the end of the exercise, students reported increased confidence in their ability to write and demonstrated higher quality scientific writing at the end of the treatment (Lee et al., 2011).

Burns, Klingbeil, and Ysseldyke (2010) studied 360 elementary schools with a mean enrollment of more than 520 students in four different states. They used a computer system that gave feedback on answers, then automatically adjusted the difficulty and type of question so that the student was forced to use that feedback on subsequent problems. Schools using this program had significant improvement over control schools as measured by performance on state standardized tests. Shute's

(2008) review of feedback has multiple other examples of learning gains and benefits attributed to feedback. The key lesson is that the evaluative feedback in the FA process improves content knowledge as well as attitudes towards science.

Since both science inquiry and FA improve content knowledge and higher order thinking, the impact should be greater when they are combined together. The idea of using FA with science inquiry is not new. Ruiz and Furtak (2007) looked at informal FA practices of 3 different middle school teachers as they taught the first four inquiry based science experiments in their classes. They found improvements in student performance and content knowledge, but did not look at data analysis skills. Likewise, Shavelson et al. (2008) found improvements in content knowledge and higher order thinking when using formal imbedded FA within an inquiry curriculum in middle school students. Carlson, Humphrey, and Reinhardt (2003) spent 12 years weaving FA into science inquiry for elementary and middle school students and found improvements in content knowledge and student performance. In each of these cases, researchers investigated the impact on content knowledge and/or higher order thinking rather than data analysis skills. Since data analysis skills and scientific explanation are two of the key practices students must learn (NRC, 2011), we have to focus on ways to improve these abilities and not just content knowledge.

In summary, formative assessment (FA) has come to mean both an instrument and a process (Dunn & Mulvenon, 2009; Bennett, 2011). While the instrument is important (Ruiz & Furtak, 2007; Shavelson, 2008; Carlson et al., 2003), educational gains are not always seen unless the instrument is used as part of a

process (Torrance, 2007; Hume & Coll, 2009). The FA process begins with assessing student understanding. Instruction is modified based on student understanding and students are given individualized feedback on how to improve their work. Students are given time to improve their work before their work is reassessed (Black & Wiliam, 1998; Hume & Coll, 2009; Ruiz & Furtak, 2007; Shavelson, 2008; Carlson et al., 2003; Shute, 2008). Of all these steps, the most important is providing students with evaluative feedback on their work (Hume & Coll, 2009; Shute, 2008). When the process is imbedded into the curriculum and evaluative feedback is stressed, students have made significant progress in their understanding of content knowledge (Ruiz & Furtak, 2007; Shavelson, 2008; Carlson et al., 2003). It is hard to determine the impact of the FA process on data analysis skills because few studies look at the impact of FA process on skills.

Assessment Creation

As was stated above, the FA process begins with assessing student understanding. To do this, appropriate assessment instruments must be used. The data analysis skills listed by the NRC (2011) include collecting data from an inquiry investigation, "analyzing data to look for patterns," making predictions and "evaluating the strength of a conclusion … based on data" (p 61-62). These align well with Leighton's (2011) classification of higher order thinking which include inquiry investigations, analyzing and interpreting data, developing inferences, predictions and arguments aligned with evidence, and substantiating/evaluating. Leighton goes on to say that higher order thinking should be evaluated using short response

questions or projects. If higher order thinking is the target of FA, the prompts in the assessment must make student thinking explicit (Furtak & Ruiz-Primo, 2008). In addition, these assessments need to be aligned to curriculum (Schafer, 2011). Combined together, this research provides a picture of the ideal chemistry FA instrument for data analysis skills. It would consist of short answer questions that make student thinking explicit while those students are using data analysis skills to evaluate chemistry content.

A search of the literature provides examples of chemistry FA instruments that focus on content (Nyachwaya, Mohamed, Roehrig, Wood, Kern, & Schneider, 2011; Doige, 2012; Branan, & Morgan, 2009; McIntosh, White, & Suter, 2009) and chemistry FA instruments that do not make student thinking explicit (Vital, 2011; Gray, Owens, Liang, & Steer, 2012). The search of the literature also found examples Math FA instruments that focused on content material and skills covered in state assessments (Phelan, Choi, Niemi, Vendlinski, Baker, & Herman, 2012; Koedinger, McLaughlin, & Heffernan, 2010), but no FA instruments that address the data analysis skills targeted by the NRC (2011).

If no ideal chemistry FA instruments focusing on data analysis skills exist, they must be created. In education programs, pre-service teachers are taught to create assessment instruments using questions from textbooks or relevant summative assessments. Unfortunately, short answer questions targeting higher order thinking are rare in chemistry textbooks. Davila and Talaquaer (2009) evaluated the review and assessment questions in a series of general chemistry textbooks. They found questions around defining a pattern, but the questions did

not include the data analysis skills targeted by this study like scientific explanation. Summative assessments do not examine targeted data analysis skills, are not tied to chemistry content, or are designed as multiple choice questions, which do not make student thinking explicit. The Trends in International Mathematics and Science Study (TIMSS, 2011) 8th grade test is designed to test analysis and reasoning as part of the test, but does not explore the specific data analysis skills targeted in this study. In addition, since they are spread across the sciences (biology, earth science, physics and chemistry), the data analysis questions were not aligned with the chemistry curriculum. Standardized multiple-choice tests were being developed to evaluate higher order thinking, but these would not make student thinking explicit (Walpuski Ropohl, & Sumfleth, 2011; Yeh 2001; Herman, 2013). If student thinking were not explicit, it would be difficult to provide the evaluative feedback that was part of the formative assessment process. This implies that the ideal chemistry FA instrument focusing on data analysis skills cannot be created using existing . textbook questions or summative assessment questions.

Since the ideal chemistry FA instrument focusing on data analysis skills cannot be created from previously validated sources, any FA instrument that is designed must be checked for validity. According to the *Standards of Educational and Psychological Testing* (1985), validity indicates how well a tests measures what it claims to measure. Cronbach and Meehl (1955) indicate there are three types of validity: criterion validity, content validity, and construct validity. Criterion validity indicates how well the instrument predicts performance. This is often used when a researcher wants to replace one test with another. Criterion validity is usually

obtained by finding a statistically significant relationship between the measure and the criterion. Content validity focuses on the content of the test and looks at whether the test is representative of the subject. Rubio devised a method to study content validity using a panel of experts (2003). The panel is selected based on each individual's expertise. Each member rates the questions validity on a scale of one through four and those responses are used to devise a content validity index. Construct validity indicates how well the assessment instrument measures a particular construct. A construct is defined as an attribute, like happiness. Construct validity is usually obtained by gathering evidence from a variety of sources including cognitive interviews, observations of behavior and correlations to existing tests that have construct validity.

Willis (2005) describes several methods of cognitive interviewing: "Think-aloud", verbal probing and other techniques. The "Think-aloud" interview involves recording a subjects cognitive stream as they go through the process of answering a question. Verbal probing is more of an investigative focus where a subject is asked a question and the researcher follows up with additional questions. Using verbal probing with "Think-aloud" techniques allows a researcher to understand the basic mental processes of a student while they are taking an assessment with follow-up into how and why certain questions were answered in certain ways.

In summary, appropriate assessment instruments must be used in order to start the formative assessment process (Leighton, 2001; Furtak & Ruiz-Primo, 2008; Schafer, 2011). To investigate data analysis skills, which are are predominantly higher order thinking, assessment instruments should consist of short answer

responses or projects (Leighton, 2011) that make student thinking explicit (Furtak & Ruiz-Primo, 2008). Assessment instruments of this type could not be found in the literature and cannot be created using existing textbook questions (Davila & Talaquaer, 2009) or summative assessments (Walpuski et al. 2011; Yeh 2001; Herman, 2013). Any new assessment instrument would have to be created using questions that have not been validated and therefore must be checked for validity (APA, 1985). Cronbach and Meehl (1955) indicate that at least content validity should be obtained for assessments looking at a skill set. Content validity can be determined using Rubio's (2003) methodology. Cognitive "Think-aloud" interviews (Willis, 2005) can be done to gain understanding of student thinking when completing assessments. Using all this research, a FA instrument could be created that gets closer to the ideal chemistry FA instrument to investigate data analysis skills.

Summary

Students struggle with data analysis skills (Hug & McNeill, 2008; Shah & Hoeffner, 2002; Kanari & Millar, 2004; Schauble et al., 1995). Many data analysis skills involve higher order thinking (NRC, 2011; Leighton, 2011). Science inquiry improves higher order thinking skills (Hofstein et al., 2004) and requires data analysis skills (Bell et al., 2005), but students need additional scaffolding in order to make use of the benefits (Sandoval, 2003; Wheeler & Bell, 2012; Keys et al.,1999).

One way to scaffold science inquiry is through evaluative feedback (Keys et al., 1999), which is part of the formative assessment (FA) process (Black & Wiliam,

1998; Hume & Coll, 2009; Shute 2008). FA has produced significant gains in student content knowledge across various contexts (Black & Wiliam, 1998; Hume & Coll, 2009; Ruiz & Furtak, 2007; Shavelson, 2008; Carlson et al., 2003; Shute, 2008), but little research has been done using formative assessment to improve data analysis skills.

The FA process begins with an assessment of student learning. The ideal chemistry FA instrument to target data analysis skills must engage and measure higher order thinking (Leighton, 2011; Ruiz-Primo & Shavelson, 1996; Furtak & Ruiz-Primo, 2008). To do this, it must consist of short response questions (Leighton, 2011) that make student thinking explicit (Furtak & Ruiz-Primo, 2008). This type of assessment does not exist and cannot be created using validated questions from existing textbooks (Davila & Talaquaer, 2009) or summative assessments (Walpuski et al. 2011; Yeh 2001; Herman, 2013). Thus, validity of the instruments needs to be measured (Cronbach & Meehl, 1955). Content validity, which is one of the three types of validity, can be measured with an expert panel (Rubio, 2003) and confirmed using cognitive interviews (Willis, 2005).

Once assessment instruments have been created, the FA process can begin. Student learning is assessed. Instruction is modified to improve student understanding. Students are given evaluative feedback on how to improve performance. At the end, student progress is reassessed (Black & Wiliam, 1998; Hume & Coll, 2009; Shute, 2008).

In this research, formative assessment was paired with science inquiry to see if evaluative feedback could improve data analysis skills in the context of a high

school chemistry classroom. Students conducted a science inquiry and feedback was given on data analysis skills. In accordance with standard formative assessment procedures, classroom instruction was modified to improve both content knowledge and skills development. Student skill development was assessed using a set of instruments created by the researcher with content validity verified using Rubio's protocol (2003). Student thinking was explored using cognitive interviews.

Methods Overview

Can the feedback associated with formative assessment be used to improve data analysis skills in high school chemistry? The goals of this mixed methods study was to investigate the following questions:

1. Can formative assessment improve data analysis skills?
2. What are student thinking processes when approaching new data analysis tasks?
3. What is the reliability and validity of the data analysis skills formative assessments that were created for this study?

To answer these questions a set of four assessments were created and reviewed by an expert panel. The panel reviewed each assessment for content validity using 11 different criteria. The panel first determined how well the scenario in each assessment accurately represented science. The panel then looked at whether the word choice was grade appropriate. Next the panel determined if the graphs were clear, and if the scenarios were clear. The panel was then asked to rate if the questions stems could effectively assess a student's ability to interpret graphs, recognize patterns, draw conclusions, test hypotheses and determine sources of error. Last the panel was asked to comment on the overall clarity of the assessment. After the panel completed their review, assessments were updated based on expert panel recommendations to improve clarity and validity.

The revised assessments were then given to two different classes. Class A was the initial treatment class. In Class A students took the pre-assessment. Class A then had instruction on data analysis skills, performed a guided science inquiry

experiment, turned in their lab report for the guided science inquiry experiment and received explicit written feedback on their report. Class A was given a chance for improving their work before being given an intermediate assessment. Two weeks later Class A was given a final assessment. Class B was given similar treatment with a small change in timing. Class B also started with the pre-assessment. Then Class B received instruction on data analysis skills, performed a guided inquiry experiment and turned in their lab report. All of these actions happened parallel to Class A. The two classes diverged when it came to the timing of the intermediate assessment and the written feedback. Instead of receiving written feedback immediately, Class B were told whether their lab reports were proficient. At this point, Class B was given the intermediate assessment. After taking the intermediate assessment, Class B was given explicit written feedback and given time to improve their work before being given the final assessment.

After both classes finished the final assessment, volunteers were solicited from both classes to participate in a cognitive interview. In the interview, students were asked to talk about their thinking processes while completing a fourth assessment that was similar to the pre-assessment, intermediate assessment, and final assessment. This type of interview is often referred to as a "Think-aloud" interview (Willis, 2005). An audio recording of the "Think-aloud" interview was made while the interviewer took notes on body language and actions. Transcripts and assessments were reviewed for insight into student approaches to new cognitive tasks.

During the course of the study, both classes covered the same content material with the same activities. The only difference was the timing of the evaluative feedback, which is crucial to formative assessment (Torrence, 2007; Hume & Coll, 2009). In the first part of the experiment, Class A served as the experimental group, receiving written evaluative feedback to guide them towards mastery. Class B served as the control group and received a statement indicating proficiency without this guidance. Between the intermediate assessment and the final assessment, Class B received their delayed evaluative feedback and were given a chance to improve.

Study design

This study was designed to assess student growth in data analysis skills due to FA. In the literature review, it was stated that the most important part of the FA process was evaluative feedback. The initial plan for the study, shown below in diagram 1, involved two classes. Two classes were selected so that one could be the experimental group, which received treatment first, and the other would be the control group, which received a delay in treatment. Both classes had an initial assessment. This aligned well with the formative assessment process outlined by Black and Wiliam (1998), which begins by assessing the current understanding of the students. Next, the experimental group received treatment, which was supposed to consist of evaluative feedback on the assessment. Both classes took an intermediate test. The control group then received evaluative feedback, which was

supposed to be on their intermediate assessment. Afterwards both classes took a final assessment.

Diagram 1: Initial diagram of study design

NO_1 X O_2 O_{34}
NO_1 O_2 X O_{34}

Key:
N – non-random sample
X – treatment + observations
O_1 – pretest
O_2 – intermediate assessment
O_{34} – post test #2 + interviews

The actual study required modification of this plan. At the beginning of the research both classes completed a data analysis skills pre-assessment. As was stated above, this aligns well with the FA process (Black& Wiliam, 1998; Hume & Coll, 2009). At that time, instruction was focused on content material. Demonstrations, questions, guided explorations and labs focused on understanding reaction rates. The pre-assessment indicated that students did not have adequate data analysis skills to continue. In addition, many of the pre-assessments were blank, which meant evaluative feedback could not be given. Part of the FA process is modification of instruction in response to student understanding. In order to abide by the process, instruction and the research design were modified as shown in Table 1.

Table 1: Revised study design

	Class A	Class B
Classes received the same treatment	Pre-assessment	Pre-assessment
	Data analysis instruction	Data analysis instruction
	Science inquiry lab	Science inquiry lab
	Science inquiry lab draft	Science inquiry lab draft
Classes received different treatment	Treatment – Feedback on science inquiry lab draft	Intermediate assessment
	Science inquiry final draft	Treatment - Feedback on science inquiry lab draft
	Intermediate assessment	Science inquiry final draft
	Final assessment	Final assessment
	Interviews	Interviews

First, additional instruction on scientific explanation was added. Leighton (2011) indicates that scientific explanations require higher order thinking. Thus, curriculum was modified using strategies that are considered best practices for teaching higher order thinking such as analyzing and relating understanding to everyday life (Miri, David & Uri, 2007). Specifically students were required to analyze scientific arguments and relate them to their lives.

Next students completed a science inquiry lab, which involved investigating the impact of different factors on the rate of a reaction. Additional data analysis instruction was added to the science inquiry lab preparation. Students were given a list of questions to help guide them in their design and prompt them to think about data analysis. In addition, students were provided a set of anchor papers to help them write their lab report and understand what appropriate data analysis looked like. Anchor papers are widely accepted as best practice and are used on the Oregon Department of Education website.

Instead of providing feedback on the assessments, evaluative feedback was provided on science inquiry lab reports. In Class A, the students were given time to revise their papers and turn them back in before taking the intermediate assessment. Class B was given an indication of whether their papers were proficient, but the students were not given evaluative feedback until after their intermediate assessment. Once Class B was given evaluative feedback they had time to revise their papers and turn them back in before taking the final assessment. Several weeks after the final assessment was given, volunteers were obtained from both classes for the "Think-aloud" cognitive interview. Diagram 2, shown on the next page, indicates how the research design functioned after the changes.

Diagram 2: Diagram of actual study

NO_1	X_1	X_2	X_3	O_2		O_3	O_4
NO_1	X_1	X_2		O_2	X_3	O_3	O_4

Key:
N – Non-random sample
X_1 – Instruction on data analysis
X_2 – Science inquiry + observations
X_3 – Treatment (evaluative feedback) + observations
O_1 – pretest
O_2 – intermediate assessment
O_3 – post test
O_4 – interviews (subset of both classes)

During the course of the study, all students in attendance were given three assessments: a pre-assessment, an intermediate assessment, and a final assessment. In each assessment, the skills being measured were the same with changes to the

context and accompanying content knowledge. The dependent variable for this study was the score of their data analysis skills as measured by the instruments and graded with a rubric. Inter-rater reliability determined using a group of three researchers. Pre-post assessment gains were measured for both groups on both the intermediate and final assessments.

Gains for the experimental group, Class A, were compared with the control group, class B, for the intermediate assessment. Gains for the control group, Class B, were compared with the experimental group, Class A, between the intermediate and final assessment. We predicted that the experimental group would have significantly higher scores on the intermediate test than the control group. It was also predicted that the overall gains would be relatively equivalent on the final assessment.

After the assessments were complete, six student volunteers were chosen for cognitive interviews. Interviewees were a mix of students from the two classes that were part of the study. The interviews were designed to be "Think-aloud" interviews conducted with the "Think-aloud" assessment to understand how students interpreted the tests as well as how different students approach new material. Following the "Think-aloud" portion of the test, the researcher continued with verbal probing into why students answered certain questions in specific ways, how students approached the problem, their understanding of the assessment and their interpretation of certain questions.

Participants

This study was performed in a public high school in the Portland metro area using 2 intact chemistry classes consisting of 10th, 11th, and 12th graders. Class A consisted of 24 participants while class B consisted of 30 participants. Of these students, we collected a full set of data from 12 students in class A and 16 from class B. Below is a table of demographics, and grade level of the students that participated from each class.

Table 2: Demographics and age of class A participants

Gender	**Ethnicity**	**Home language**	**Grade**	**age**
71% male	46% White	67% English	54% 10th grade	8% 15
29% female	29% Asian	13% Vietnamese	46% 11 & 12 grade	58% 16
	13% Black	8% Spanish		33% 17 & older
	13% Hispan	13% other		

Table 3: Demographics and age of class B participants

Gender	**Ethnicity**	**Home language**	**Grade**	**age**
63% male	23% White	53% English	67% 10th grade	32% 15
37% female	27% Asian	10% Vietnamese	33% 11th grade	53% 16
	17% Black	23% Spanish		23% 17 & older
	20% Hispani	13% other		
	13% Multi			

Treatment:

Both classes that participated in the study were provided the first assessment during the same week. It was the original intent that students would receive feedback on this assessment, but approximately 50% of the assessments were returned either blank or with statements indicating that the students didn't know how to answer the questions. As was mentioned above, it was determined

that additional instruction was needed prior to the second assessment. In addition, students would be unable to receive feedback on their pre-assessment. Instead, feedback would have to be given on the science inquiry lab report.

A review of the completed and partially completed assessments indicated that many students had difficulty with the following areas: graph interpretation (which should have been called scientific explanation), error analysis, and hypothesis testing. The corresponding Hug and McNeill (2008) categories are drawing conclusions, limitations of data, and use of content knowledge. Drawing conclusions most closely relates to the NRC (2011) concept of scientific argumentation and what much of the literature refers to as scientific explanation (Ruiz-Primo, Li, Tsai, & Schneider, 2010; Sandoval, 2003; McNeill, & Krajcik, 2007; Keiler, 2007; McNeill, Lizotte, Krajcik, & Marx, 2006). Scientific explanation consists of three parts: claim, evidence and reasoning (Ruiz-Primo et al., 2010). Ruiz-Primo et al. (2010) define claim as "a testable statement or conclusion that answers a scientific question" (p. 4). Evidence includes data used to "construct, support and defend a claim" (p. 4). Reasoning connects the evidence to the claim through "a conceptual and theoretical link" (p. 4). Together, the three parts should be a cohesive unit that allows a student to demonstrate understanding and thinking processes.

How should scientific explanation skills be taught? Driver et al. (2000) indicates that scientific explanation skills must be worked on in a social setting such as a class discussion. The Sandoval (2003) found that scaffolding helped students construct scientific explanations. Sandoval's scaffolding consisted of helping

students make a claim and linking that claim to data. Based on these two recommendations, an intervention was planned that would scaffold the creation of scientific explanation in a group setting. Students were given arguments for and against water Fluoridation (Appendix A). This topic was chosen because it was subject of current debate, in the news and directly affected student's lives. Students analyzed arguments and were told to find the three parts of the argument. Students were instructed to color code claim statements with one color, reasoning statements with another color and evidence statements with a third color. Students then summarized the argument for and against water quality together as a class. The goal of this part was to help them see how the data was tied to the conclusions through reasoning. Afterwards in the same class session, students constructed an argument for who was the best musician using claim, evidence and reasoning. This took the idea of constructing an argument and related it to a familiar non-science setting that was part of the student's lives, which is one of the best practices listed for teaching higher order thinking (Miri, David & Uri, 2007).

This foundation of scientific explanation was built upon through a science inquiry project. Students learned about reaction rates in the classroom and were given the choice of designing an inquiry project around one of the science demos we had used: iodine clock reaction, Mentos and Diet Coke or Alka-Seltzer and water. Students were told to design an experiment that involved altering one the different factors that was discussed in class: concentration, catalyst/inhibitor, temperature, and surface area. Each of these factors impacted reaction rate in some way.

Students were told to design their own experiments, but most students did not know where to begin. Students were given more freedom on this experiment than they had been given in previously. Although parts of experimental design had been discussed, most students walked into the lab without any plan. In response to the student's lack of planning, individual group meetings were held to flesh out experimental design and testing methods. Students were told to have at least three tests so they would have at least enough data to make a decision and advice was given to several groups on how to test the factors they were interested in.

The three problem areas that showed up on the completed pre-assessment, graph interpretation (which should have been called scientific explanation), hypothesis testing and limitations of data, were addressed in the packet they were given to assist with their lab report write up. Students were given a sheet of questions to help them write their lab report (Appendix B). In addition, they were given two example papers (Appendix B). One was an example of a lab report that was proficient and one was not proficient. These were written based on an expository lab the students had completed earlier in the unit. Students were given a simplified Oregon Department of Education scoring rubric (Appendix B) and asked to determine what made one paper proficient and the other non-proficient. A class discussion was held where the specific issues with hypothesis formation, scientific explanation and limitations with data were discussed.

The goal of this packet was to help students understand how to do the data analysis portions of the lab assignment and understand the expectations of the assignment. Part of the formative assessment process is educating students about

expectations. The evaluative feedback that was the focus of this study is intended to help students see the gap between their work and the expectations. In order for the students to understand and internalize the feedback, they had to be given an opportunity to understand the expectations first.

Only one student from both classes combined created a proficient report based on the models and the rubric given. Students in Class A were given explicit written feedback on their lab report. This feedback tied their work to expectations in the ODE rubric so they would understand what was needed in order to create a proficient paper. Students in Class A were given time to read through their feedback and class time was taken to explain the most common comments that were given. Examples from both student work and from the model papers were used to illustrate how to make final lab reports proficient. Students were given time to ask questions about their specific feedback and work on their papers. They were asked to rewrite the papers to make them proficient and told that papers would only be accepted for credit if they were proficient. The class period after students were supposed to turn in their revised papers, Class A received the intermediate assessment.

Class B received their lab reports with a "P" or "NP" mark indicating whether they were proficient or not. While in some classrooms this kind of mark could be considered a grade, in this classroom a non-proficient mark indicates that more work is needed. Class B also received an additional copy of the scoring rubric. They were given time in class to ask any questions they wished about their papers and told they would need to rewrite their papers to make them proficient. A few

students asked questions about the papers and these were answered. Class B was then given the intermediate assessment. One week after they were given the assessment, Class B was given explicit written feedback on their papers. They also received explanation of the most common issues that were seen with examples from the two papers model papers and from the class. They were given class time to ask questions about the feedback they received and reminded that they needed to rewrite their papers. Just like Class A, they were told that they would only receive credit for the papers if they were proficient. After Class B turned in their final draft papers, both classes were given the final assessment.

The goal in splitting the treatment over two classes was to be able to quantify gains that were specifically due to the feedback given as part of FA. The goal was to screen out other potential factors such as teacher instruction and content knowledge. The two classes chosen were picked because they were the most similar classes in terms of size and ability that were available for study. Classes available for study were only those in which the researcher was student teaching.

Instruments

Several data analysis skills were highlighted for this study: graph interpretation, recognizing patterns/making inferences and drawing conclusions from data. Graph interpretation was defined according to Hug and McNeill (2008) as the ability to read and interpret a graph. The skill of recognizing patterns and making inferences involves the ability to observe the graphs they have produced and infer ideas from the data. For example, if students plot the concentration of two

things ocean acidity and percent carbon dioxide dissolved versus time, they should be able to infer that the dissolved carbon dioxide was affecting the ocean acidity. Drawing conclusions was aligned with scientific explanation and defined by Ruiz Primo et al. (2010) as a claim, supported with relevant evidence and solid reasoning.

A literature search was performed in an attempt to find existing instruments to assess these skills. None were found. It was determined that standard textbook questions were inappropriate (Davila & Talaquaer, 2009), and standardized tests did not adequately assess these higher order skills (Walpuski et al. 2011, Mullis 2009, Wygoda & Teague 1995, Yeh, 2001). In addition, because standardized tests have a high percentage of multiple-choice questions, they do not make thinking explicit. Furtak and Ruiz-Primo (2008) indicate that it is essential to make student thinking explicit in assessments if those assessments are to be used in a formative manner. Tests such as the smarter balanced assessments that are being designed to look at these skills were designed for math assessment and were not available to the researcher for use.

A set of four assessments was created. Three were to be given to all students and the last was to be used as part of a "Think-aloud" cognitive interview. Each assessment was constructed in a similar fashion starting with a science scenario, which consisted of an explanation, a diagram, and an accompanying graph. Difficult terms were defined at the bottom. Each came with one page of questions, which are discussed below. For each assessment, the scenario was different and intended to follow the course of material being taught in the classroom.

It should be noted, that after the assessments were created and checked by the expert panel, the order of subject material being taught in the classroom changed. As a result, students were studying reaction rates during most of the research instead of electrochemistry. This meant that the intermediate assessment was not as connected to what students were learning at the time.

The questions were parallel for all assessments. The first question asked the students to identify a certain point on the graph (graph interpretation). Students were then instructed to explain in their own words what was happening in the graph (scientific explanation). Students were prompted to include a claim, evidence, and reasoning in their explanation. The third set of questions asked them to predict based on the trend what would happen on a point that was not on the graph. The student was instructed to assume that the trend would continue and explain why they thought it would happen or not (inferences and patterns).

The next question was intended to address hypothesis testing. Students were presented with a claim made by a friend or lab partner. This claim related to the scenario they were given. Students were asked to provide what kinds of data that would need to prove that the claim was wrong. The last question was intended to address error analysis. Students were asked to list data they would need in order to determine if the measurements were accurate.

Assessments were graded on a standardized rubric shown in Appendix C. Below is a summary of the topics that were used for the scenario in each of the assessments.

Pretest:

The pre-assessment focused on ocean acidity. Students were given a scenario that explained how water was being tested for dissolved carbon dioxide and acidity at a station in Hawaii. The scenario also described how dissolved carbon dioxide turned into carbonic acid. Students were given an accompanying graph that showed the measurements of dissolved carbon dioxide and surface level ocean acidity.

This topic was chosen because students had just completed a unit on acids and bases. They should have had familiarity with what constituted an acid and a base as well as the chemical reactions. The biology teachers that students were exposed to in their freshman year taught lessons on global warming. Conversations with students indicated that many had been exposed to the idea through media. In addition, several students were part of Ocean Bowl, which covers ocean acidification as one of the topics. The full assessment as it was given to the students is in Appendix D.

Intermediate test:

The intermediate test focused on plating a silver spoon. Students were given a graphic indicating a plating setup and a graph indicating measurements of mass at different times for the anode and cathode. The paragraph gave a simplified overview of plating with new words defined.

This topic was chosen because students should have been halfway through their electrochemistry unit. The way that the unit was originally planned, students should have learned about oxidation and reduction as well as basic ideas about what

happens at the anode and cathode of a cell. As was stated before, the order of topic was changed after assessments were created. Instead of learning about oxidation and reduction students were studying reaction rates. As a result, this assessment was not aligned with the curriculum as was originally intended. The full assessment as it was given to the students is in Appendix D.

Post-test:

The post assessment included a graph about rechargeable batteries. It plotted the battery capacity, self-discharge rate and internal resistance versus the number of times that the batteries were recharged. The paragraph described how batteries changed with charging and compared them to cell phone batteries.

This topic was chosen because students should have been completing their electrochemistry unit at this time. Students were supposed to create a zinc-copper electrochemical cell and used this to power a calculator. They were supposed to have exposure to a video on how batteries work and be able to relate this to the cell they created. This assessment extended the topic further by showing them more information about battery charging and relating this to things they knew.

Instead, students received this assessment as they were just beginning electrochemistry. They had learned about oxidation and reduction, but had not yet made their zinc-copper cell. They understood some of the background, but once again this assessment was not as well aligned with the curriculum. The full assessment as it was given to the students is in Appendix D.

“Think-aloud” assessment:

The “Think-aloud” assessment was more complicated than any of the other assessments. It gave a diagram of a UV-VIS spectrometer and a picture of nanoparticles. The graph showed the increasing absorbance with time as nanoparticles were created. The accompanying paragraph described how a fungus was used to grow silver nanoparticles and that the number of particles grown corresponded to the increasing absorbance.

This topic was chosen because students were supposed to have exposure to nanomaterials through a set of NOVA videos that were supposed to be shown. They were to look at how electron transfer related to photochemistry and light. Due to time constraints, this video was not shown. Thus, this assessment was not aligned with the curriculum. The full assessment as it was given to the students is in Appendix D.

Interviews

When the final assessment was given, volunteers were solicited from both classes to participate in a cognitive interview. Students were informed that interviewees would be asked to complete an additional assessment and provide feedback. Multiple students volunteered and six were chosen. Students that were chosen fell into three different categories: high performing, average performing and low performing. High performing students were defined as students who scored highly on at least one assessment and who had a high grade (A or B) before this unit started. Average performing students had moderate scores on at least one

assessment or had an average grade (B or C) before the reaction rate unit was started. Low performing students had poor scores on all assessments and low grades in the class (D or F). They were students who constantly retook tests in efforts to pass, but many were still failing at the end of the term. There were two interviewees that fell into each category.

The goal of the cognitive interview was to help understand student thinking during these assessments. The intention was to see how students of different abilities approached the task. Students were chosen from both classes because at the end of the study both classes should have been matched.

Interviews were done in a quiet room and were audio recorded. During the interview, students were asked to explain their thinking while completing an assessment. Standard cognitive interview protocols were observed (Willis, 2005). Students were given a warm up exercise where they figured out how many windows were in their house while explaining their thinking. Students were then asked to complete the "Think-aloud" assessment found in Appendix D while talking about their thinking process. During this time, the researcher took notes about student movements and actions that indicted strategies that were not being discussed, such as the student referring back to passages in the paragraph.

There was a semi-structured protocol to the portion of the interviews that followed the "Think-aloud." Table 4 below shows which items were planned and a list of common questions that were added after the interview was over.

Table 4: interview questions

Planned questions	Additional questions
What did you think these questions were asking? (referring to specific questions) How do you approach questions when you do not understand the material? Did you feel like you understood this scenario?	How could I rephrase the last question to better ask what were the sources of error? Did you feel like you learned while you were taking this assessment? You went back and re-read sections. Why did you do that? What were your goals? Do you understand the scenario more now than when you first read the assessment?

These interviews allowed the researcher to gain insight into the learning process and student thinking. Questions asked helped highlight issues with the assessments as well as ways that students interpret questions. Follow-up was intended for improving the assessments for future use as well as understanding the ways that students responded on the other three assessments.

Expert panel review

A number of experts were selected for their experience in chemistry, teaching and/or education research background. All were contacted via email with a cover letter indicating the nature of the study. The letter indicated both the scope of the research and the desired actions of the experts. The final panel consisted of three expert teachers, three PhD chemists, one professor of education, and one person who had both a PhD in chemistry and a PhD in education.

Each expert was referred to a confidential on-line survey that included detailed questions about each aspect of each assessment. The 84-question survey asked the experts to rate each question, graphic and scenario on a four-point scale for language, science content and age appropriateness. Ratings were done similarly to Rubio (2003) with one indicating unacceptable, two indicating needing major revisions, three indicating needing minor revisions and four indicating that the section was acceptable. For every section that needed revisions, open text boxes were provided to allow the experts to comment on the nature and types of revisions that were needed to ensure that the questions were clear and appropriate.

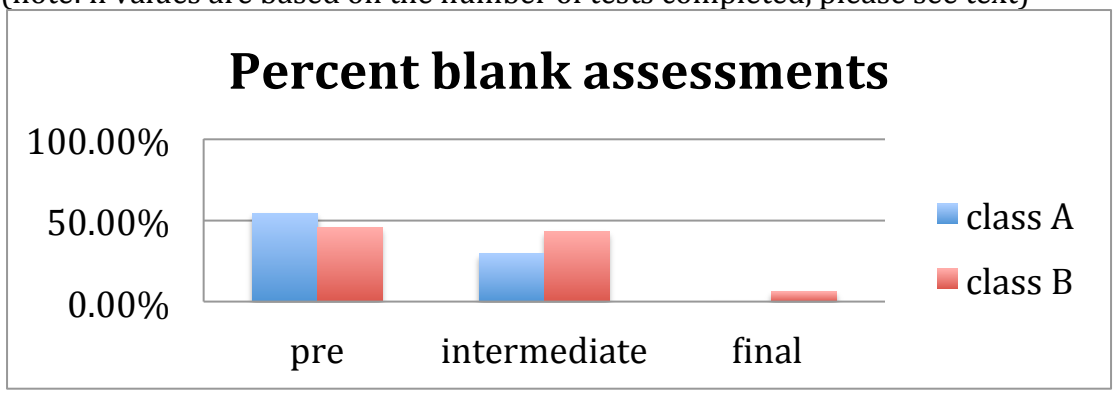
Content validity was analyzed using Rubio's content validity index (CVI) from his 2003 paper. CVI for each question was calculated by counting the number of experts who rated a question as a 3 or 4 and dividing by the number of experts. Any questions that did not receive a CVI score of 0.8 or greater were revised according to feedback. In addition, feedback was reviewed and incorporated as appropriate into the remaining questions on the assessments before they were given to students. CVI for the measure was taken as the average CVI of all the questions in the measure.

Results

Impact of formative assessment on data skills

There were a large percentage of blank pre-assessments or assessments filled out with just "I don't know." Results are summarized in Figure 1. Due to changes in attendance for each day, n values varied for each test. In Class A, 22 pre-assessments were given but 12 or 54% of these were blank. There were 24 pre-assessments given in class B and 11 or 46% of these were blank. This percentage trended down with subsequent assessments. In Class A, only five of the seventeen intermediate tests were returned blank or 29% whereas in Class B ten out of the 23 intermediate tests were blank or 43%. (note: Some of the improvement in Class A could be attributed to absences, as three of the individuals who turned in blank pre-assessments were not present for the intermediate assessment. If we remove those individuals from the numbers, Class A still had 48% blank pre-assessments compared to 29% blank intermediate assessments.) To relate this back to the study, the class that received the treatment prior to the assessment, Class A, had the blank response rate drop to 29% whereas in the class that received a delay in treatment, Class B, 43% of the assessments remained blank. On the final assessment, the percentage of tests that were blank dropped even further. In Class A, none of the fourteen assessments were completely blank, but two respondents left the last question blank. This was recorded as having 0% blank assessments for Class A. Class B had only one blank assessment out of sixteen, which represents 6% of the class.

Figure 1: Percentages of blank assessments
(note: n values are based on the number of tests completed, please see text)



For inter-rater reliability, the intermediate assessments were scored with the scoring rubric provided in Appendix C. Inter-rater reliability was calculated in SPSS for intra-class correlation using a 2-way model. Each question group was calculated individually and scored for absolute agreement on a single measure. Alpha values for the IRR scores are listed in Table 5 below. Questions 1, 2 and 4 had an alpha value of 0.7 or higher, which is considered acceptable. Question 1 was graph comprehension. Question 2 was the set of questions around graph interpretation, which should have been labeled scientific explanation, and question 4 was hypothesis testing. Those questions where IRR was not deemed acceptable were questions 3, which were the set of questions around inferences and patterns, and question 5, which was the question around error analysis.

Table 5. IRR coefficients for the scoring of the questions against the rubric for the intermediate assessment

Question set	1	2	3	4	5
IRR score	0.86	0.74	0.51	0.77	0.35

To determine learning gains, all assessments were graded using the rubric in Appendix C. Graph comprehension was graded on a scale of 0-2 with 0 being irrelevant data or blank and 2 being a correct number based on the graph. Graph Interpretation (scientific explanation) was a composite score summed from the three parts of explanation: claim, evidence, and reasoning. Each of these subcomponents was graded on a scale of 0-2. As a result, scores for graph interpretation could vary from 0-6. The inference and pattern consisted of two parts. The first asked students to predict a point not on the graph using the data that existed and the second part asked them to explain their prediction based on the patterns they saw in the data. The scores were a sum of those two components and could vary from 0-4. The hypothesis testing was scored on a scale of 0-2 and the error analysis was scored on a scale from 0-4.

There were two classes used for this study. As was stated before, both classes received modified instruction after the pre-assessment and participated in the science inquiry lab. The difference between the classes was predominantly the timing of treatment, or evaluative feedback on their science inquiry lab report. Class A received the treatment between the pre-assessment and the intermediate assessment. Class B received the treatment between the intermediate assessment and the final assessment. Data analysis skills gains were determined in two ways. First classes were compared with themselves to determine if statistically significant learning gains were achieved, Table 6.

Table 6: Paired t-test results for classes vs. pre-assessment

	Intermediate delta						Post assessment delta				
	Graph comprehension	Graph interpretation	Inferences & pattern	Hypothesis testing	Error analysis		Graph comprehension	Graph interpretation	Inferences & patterns	Hypothesis testing	Error analysis
Class A n=12	0.23	0.001 *	0.37	<0.001 *	0.22		0.996	<0.001 *	0.17	0.001 *	0.1
Class B n=16	0.06	0.3	1	0.04*	0.17		0.14	0.1	0.42	0.003 *	0.05

Note: statistically significant differences are noted with a *

A paired t-test was used to compare student scores on the intermediate to the pre-assessment and to compare final scores to the pre-assessment. This was used to determine if any skills were attained. The second comparison looked at the differences between the two classes to determine if the treatment had any affect. A t-test was run for scores on the intermediate assessment between Class A and Class B. Additionally, there was interest in whether the two classes ended up matched at the end of the study. Final assessment scores for Class A and Class B were compared by standard t-test. The p-values for the t-tests comparing the classes are shown in Table 7.

Table 7: T-test results comparing class A to class B for the intermediate and final assessments

Intermediate deltas	Final deltas

	Graph comprehension	Graph interpretation	Inferences & patterns	Hypothesis testing	Error analysis	Graph comprehension	Graph interpretation	Inferences & patterns	Hypothesis testing	Error analysis
Class A vs. B	0.07	0.01*	0.05	<0.001*	0.43	0.001*	0.03*	0.25	0.07	0.50

Note: statistically significant differences are noted with a *

Student learning – comparison to pre-assessment: If student scores on the intermediate test are compared to the pretest, Class A shows statistical differences for graph interpretation and hypothesis testing. These scores indicate that students did understand more about these two data analysis skills after the combination of modified instruction, science inquiry and evaluative feedback. Class B showed statistical gains in only hypothesis testing after participating in the modified instruction and the science inquiry. When comparing the final scores to the pre-assessment, this trend does not change. No changes were seen in the scores for inferences and patterns and error analysis. This may be related to scoring issues and will be reviewed in the discussion section.

Student learning – comparison between two classes: When the two classes are compared to each other, there are statistical differences between the classes with respect to graph interpretation (scientific explanation) and hypothesis testing. Class A, which received evaluative feedback on their science inquiry lab report before the intermediate test, performed statistically better in these two areas. At the end of the study, there was still a difference in graph interpretation (scientific

explanation) but the difference between the two classes in hypothesis testing had gone away. A new difference showed up between the two classes for graph comprehension. On the assessment, this question asked students to identify a specific point on the graph. Class A showed a decline in this area whereas Class B showed a small increase. Further exploration is needed to understand this difference.

Student learning by demographic data: When scores were examined with respect to gender, there was no statistical difference. When split by ethnicity, there was only one statistical difference noted between groups. Multiethnic students had smaller gains than white students, but this was driven predominantly by the small sample of multiethnic students. There were no statistical differences in overall skill gains when data was split by first language.

In summary, there were overall learning gains for graph interpretation (scientific explanation) and hypothesis testing for the class that received the treatment first, Class A. The students improved both their overall scores as well as their scores in comparison to the other class on the intermediate test. Scores on the final test were not matched between the classes indicating that all of the learning gains may not be attributable to evaluative feedback alone. Further research needs to be done.

Student thinking processes – Science inquiry lab

Students from both classes performed poorly on the initial draft of the science inquiry lab report. Of the 54 students in the two classes combined, 45 initial drafts were turned in and only one was considered proficient as graded by the ODE

rubric. Students were given evaluative feedback on their science inquiry lab reports and all but one of the final drafts that were turned in showed some improvement. Unfortunately, many did not have enough improvement to make them proficient. In fact, of the final drafts that were turned in, only eight were considered proficient.

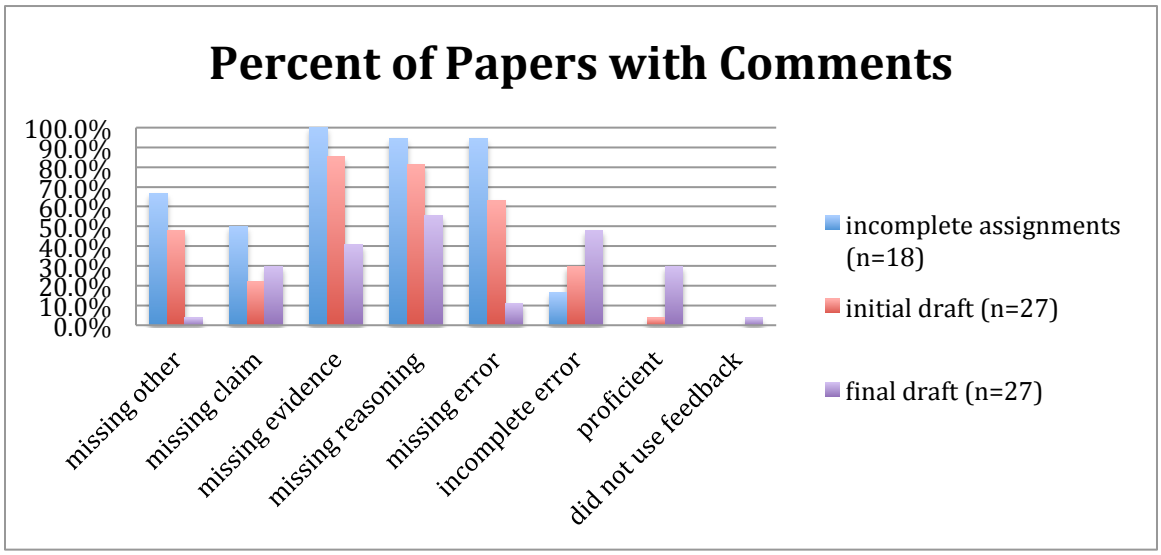
On each initial draft and final draft turned in, students were given directed feedback based on the targeted goals listed in the simplified ODE rubric in Appendix B. While data analysis was not stated in the rubric, expectations were laid out with respect to including error analysis. Feedback was given on error analysis as well. The types of comments given fell into six categories: missing other sections, missing claim, missing evidence, missing reasoning, missing error analysis and incomplete error analysis. Missing other sections refers to sections of the paper that were missing but not specifically linked to the goals of this study. This includes missing a hypothesis, failure to write out procedures and failure to record data. Claims such as "My hypothesis was right" without any further statement were considered to be missing, as were blanks. In addition, incomprehensible or incomplete statements like "took the most" without out reference to what took the most were also considered to be missing a claim. Evidence required two pieces of data that supported the claim and reasoning required students to apply what they learned about reaction rates to explain why a certain reaction went faster or slower. Finally, sources of error were considered complete if they listed at least three of the four types of error: design, measurement, execution, and representation.

The initial intention of classifying comments into these categories was to compare comment frequency on initial science inquiry lab reports versus final

drafts of those same reports. Lack of student participation interfered with this analysis. Initially 45 papers were received from the 54 students in the class, but only 27 final drafts were turned in. Eighteen students turned in initial drafts, but never revised their papers to turn in final drafts. As a result, it was difficult to compare comment frequency on initial drafts versus final drafts. To alleviate this issue, the science inquiry lab reports were split into two groups. The first group, which consisted of 27 papers, included all science inquiry lab reports where an initial and final draft were scored. On these papers, the frequency of comments was obtained for both the initial and final papers. These comment frequencies were converted to percentages so they could be compared to the frequency of comments on incomplete assignments, where only an initial draft was turned in. Comment frequency percentages on both the initial and final drafts of the science inquiry lab reports are shown in Figure 2.

The second group of papers consisted of the 18 initial draft papers that were not revised. These papers were considered incomplete assignments because students did not finish revisions before the end of the school year. The frequency of comments on this group of papers was determined and converted to a percentage so they could be compared to the frequency of comments on the first group of science inquiry lab reports. This second group of papers is also shown in Figure 2.

Figure 2: Feedback made on initial and final drafts of science inquiry lab paper



In the first group of papers, those that were revised, over 80% of initial drafts had issues with scientific explanation (claim, evidence and/or reasoning) and over 90% had issues with error analysis (either missing error analysis completely or incomplete error analysis). In addition, 48% of the papers had other incomplete sections like a missing or incomplete hypothesis, missing or incomplete procedures or missing or incomplete data table. On the final draft of these papers, Figure 2 shows that the percentage of papers that were missing data analysis pieces like claim, evidence, reasoning, and error analysis decreased to 55% or below. The number of papers with incomplete error analysis increased from 30% to 48%, but the number of papers with error analysis issues dropped to less than 60%. In comparing initial drafts to final drafts, all but one paper incorporated feedback, making papers closer to proficient.

The eighteen initial drafts that were considered part of the incomplete assignment group were given a higher percentage of negative comments. All of these papers were missing evidence and had issues with error analysis. Almost 95% were missing reasoning as well and 67% were missing other sections of the report. The percentage of papers missing claims was double those in the other group, almost 50%. Since final drafts were not received on these papers, the students were considered to have avoided the task. Potential reasons for these missing assignments are given in the discussion section of this paper.

In summary, initial scores on the science inquiry lab report were low. Only one paper was proficient. Students demonstrated one of three behaviors with respect to the lab report: resolving issues, attempting to resolve issues, and avoiding the challenge. Those that resolved the issues understood and incorporated all their feedback. This accounted for approximately 15% of the students involved in the study. Those that attempted to resolve issues incorporated some of the feedback. This group accounts for approximately 35% of the students involved in the study. This study did not explore why these students did not incorporate all of the feedback they were given. The remainder of the students either didn't complete an initial draft or did not revise the initial draft of the paper. These students are classified as avoiding the challenge and account for 50% of the students involved in the study. Although students who participated in the work improved, half of the students were disengaged or avoided the science inquiry lab report.

Science inquiry lab observations

Although the cooperating teacher indicated that students had experience with science inquiry, student behaviors indicated that this task was new. Students were allowed to pick a reaction they were interested in and were clustered into groups by the reaction they chose. The options were Mentos and Diet Coke, Alka-Seltzer and water, or the iodine clock reaction. A small minority of students, those who had done work on science fair projects, began work immediately and did not need instruction. The remainder of the class had difficulty explaining what they were testing and how. It became clear on the day of the lab that students needed more scaffolding.

To help these struggling students, the researcher asked the students to recall the different factors had been discussed in class with respect to reaction rate: temperature, concentration, catalyst/inhibitor, and surface area. They could come up with the list, but struggled with how to use that list of factors to design an experiment. These struggling students were put together into groups based on the reaction they chose and partnered with either the cooperating teacher or the researcher. The groups discussed what factors could be tested, how those factors could be tested and what their data table might look like. Many students created acceptable experimental plans, outlined data tables and began their experiments at this point.

The remaining students talked and gossiped until the cooperating teacher or the researcher followed up with these groups. When confronted, this last group of students indicated that they did not understand what they were supposed to do. The

researcher walked them through the process of designing an experiment. Students were asked to identify what factor they wanted to test and specify three levels of this factor. For example, if the student wanted to test temperature, they were asked to describe how they would create three different temperatures for the reaction. Most of these students did not know how to answer the question even though a hot plate was on the counter as was a container of ice. These students did not indicate understanding until the researcher stated they could make hot water on the hot plate and use the ice to make the water cold. Afterwards they had to be told to use the thermometers to record the temperature of the water. This interaction transformed the lab experience for these students from guided inquiry, where only the question is given, into structured inquiry because the question and procedures were provided (Wheeler & Bell, 2012).

Many of these same students were also confused about how to do the lab write-up. Before doing the write up, the assignment was discussed as a class. Students were given two example papers and told to determine what was good and bad about each. Even though this was done, students did not seem to know how to write a hypothesis or how to structure their results. Questions also arose about how long the paper should be and how to do the analysis. Questions were answered and students were given a class period to work on their lab report. Scoring on the science inquiry lab reports, discussed above, indicates that many of these students still did not understand what was expected or how to perform these tasks.

The science inquiry lab highlighted various behaviors for dealing with new tasks. They included engagement, questioning and avoidance. Students familiar with

the task were quickly engaged in their laboratory experiment. Those who were uncertain about the task asked questions until they felt they could complete it. The students that avoided the task did not participate until the task was altered to make it structured inquiry, which was more familiar to them. When this was done, the cognitive load of the task was dropped reducing the amount of higher order thinking required. Implications of this are reviewed in the discussion section.

Student thinking processes – cognitive interviews

Cognitive interviews were done using volunteers from both Class A and Class B. The interviewees spanned the spectrum of different abilities within the two classes. Students were classified into three groups (higher achievers, average achievers and low achievers) with two students coming from each group. These designations were made based on student assessment results on the final assessment as well as average scores within the class. The high achieving students were defined as those that had an A or B in the class. They usually passed exams on the first try. They were often diligent about their work when in class and required a small list of minor revisions to make their science inquiry paper proficient. They incorporated all or most of the feedback given on their science inquiry report and turned in proficient or nearly proficient science inquiry lab reports. Their final assessments included long detailed descriptions and they scored above average on most questions. Average achieving students had a B or a C in the class prior to this unit of instruction. They had average scores on the final assessment and required a long list of revisions to make their science inquiry paper proficient. These students

turned in final drafts of their science inquiry lab reports that were not proficient, but they did incorporate much of the feedback that was given. Low achieving students were defined as those that struggled to obtain proficiency in the class and had to retake multiple exams in order to pass. They performed poorly on their final assessment and required major revisions on their science inquiry paper in order to be proficient. In many cases, their final version of the science inquiry paper was still not proficient when graded using the ODE rubric or no final paper was turned in.

Cognitive interviews were done with the "Think-aloud" assessment in Appendix D. This assessment had an average CVI score of 0.78, which was below the threshold of 0.8 set in the Rubio paper (2003). After the expert panel reviewed this assessment feedback was incorporated to improve the assessment, but there are indications listed below that indicate it was did not have content validity. Due to time constraints, the expert panel was unable to review the final version of the "Think-aloud" assessment to determine if the CVI score improved.

Student behaviors varied by group. High achieving students read the scenario carefully going back several times to try to relate the material to the graph and the questions. Even after this was done, they indicated that they did not fully understand the assessment. They appeared to learn while they were taking the assessment with their answers indicating more understanding as they moved through the assessment, but they did not feel confident in their understanding at the end. After completing the assessment, the high achieving students indicated that the assessment instrument did not have enough information to give reasoning for the scientific explanation and did not have enough information to answer either the

hypothesis testing question or the error analysis question. They noted that the "Think-aloud" assessment was more difficult than the other assessments they had previously been given.

Average achieving students approached the assessment differently. They both tried answering as many questions as possible using only the diagrams and pictures before reading the material. In one case, the student commented that it was too much reading. The other student tried to go back and read through the material, but struggled. She attempted to use the vocabulary to help her, but found that it did not. This is talked about further in the discussion section.

Low achieving students struggled with comprehension. One struggled reading the words, but kept circling back through the vocabulary sounding out ideas and trying to make meaning of the context. He continually referenced back to the scenario in trying to make meaning of things and underlining passages. In many ways, his strategies resembled those of the high level learners, but his performance on the assessment indicated low levels of comprehension.

The other student in the low achieving group appeared to be working as fast as possible to complete the assessment. He would half read sentences and pause in the middle saying "oh" as though it suddenly made sense. Many of his answers seemed to be just a shuffling around of vocabulary terms. The second to last question asked "Your partner says he has some fungus in his back yard that could make silver nanoparticles. What data would you need to prove your partner wrong?" He stated out loud multiple times that he needed to know what kind of fungus was being used, but chose not to write that answer down on the paper. After the

assessment was complete, the researcher asked him why he didn't write down his question about the fungus. He indicated that he interpreted the question to be about proving his friend wrong and missed the part about what kind of data he would need.

In spite of the differences in approach to the assessment instrument, there were commonalities in the outcomes. Students from all groups struggled with developing the reasoning part of their scientific explanation. High achieving students indicated the assessment didn't have enough information. Average achieving students used the graph as evidence and low achieving students created a mix of words from the scenario that did not make sense such as "as the wavelength goes up the absorbance goes up with hours."

Students in all groups seemed to be comfortable with extrapolating from the graph, even when they didn't understand what was happening in the scenario. When asked to defend their claims, the most students answered, "because it does" or "I think it will." The intention from this group of questions was to have students explore inferences and patterns, but students did not read the question in this fashion.

Students from all groups misinterpreted the last two questions as well. The second to last question was intended to ask about hypothesis testing, but students focused on how to prove someone wrong instead. When the students were questioned about this, many said they didn't really have enough information to test the hypothesis. The last question was intended to ask about error analysis. The answers given indicated students did not interpret the question in this way. One

student responded “The same” and drew an arrow to the previous question. Another student indicated she didn’t know what the question meant. In the verbal probing after the “Think-aloud” interview several students confirmed that they interpreted the question to be about their friend’s experiment instead of the whole scenario.

After the “Think-aloud” interview, the pre-assessment, intermediate assessment and final assessment were reviewed again. There was evidence on those assessments that many students misinterpreted questions on the interviews. A list of common misconceptions is show in Table 8 below.

While multiple issues were highlighted with the “Think-aloud” assessment, it still provided insight into student thinking on new tasks. Higher achieving students used literacy strategies to help them answer the questions. Average and low achieving students struggled with comprehension and may have benefitted from the use of literacy strategies. That lack of reading comprehension hampered average and low achieving student’s ability to perform the data analysis tasks. These issues with comprehension were highlighted more in the verbal probing part of the cognitive interview. All students indicated they did not really understand the assessment. Most of the average and low achieving students indicated they did not have strategies to read scientific writing.

Table 8: Common misinterpretations of assessment questions

Question	Intended response	Misinterpretation
In your own words describe what is happening in the graph	Students would read the scenario as background for the graph and use this to help explain what was happening	Explain that lines were going up or down without referencing the scenario. Could not give any reasoning
What data would you need to prove your partner wrong?	Students would plan a way to test their friend's hypothesis	Students indicated they would need more data provided by someone else without indicating what that data was. Students focused on proving their friend right or wrong without talking about the data. "He is right."
What data would you need to determine if these measurements were accurate?	Students would look for sources of error in the scenario	Students indicated this was a redundant question. "The same!" or answer copied from previous question Students interpreted this to look for sources of error in their friend's hypothesis.

Assessment validity

Many of the points that were highlighted in the cognitive interviews point to issues with assessment validity. Table 9 contains the content validity evaluations for each question based on the feedback from the expert panel.

Table 9: CVI scores from expert panel review of question content validity

Question	Pre-assessment	Intermediate assessment	Final assessment	Think aloud
Accurately represent science	1.00	1.00	1.00	0.88
Scenario word choice	0.88	0.88	0.75	0.63
Graph clear	1.00	0.88	1.00	0.75
Scenario clear	1.00	1.00	0.88	0.75
Graph interpretation	0.88	0.88	1.00	0.75
Recognizing patterns	1.00	1.00	0.88	1.00
Draw conclusions	1.00	1.00	0.88	0.75
Test hypothesis	0.88	0.75	0.75	0.75
Sources of error	0.88	0.88	0.63	0.75
Overall clarity	1.00	1.00	1.00	0.78

On each of the first three assessments (pre-assessment, intermediate assessment and post assessment) a majority of the questions received a CVI score of >0.8. Three questions, one from each assessment, received a CVI of 0.78. Each of these was modified according to the expert panel comments as listed below.

On the pre-assessment the experts had issues with the question on testing hypotheses. Students were asked to plan how they would test the water pH in 2015. Many experts felt like this was too far into the future. The time frame was changed to August of 2013 to make the question time frame more reasonable. On the intermediate assessment, the experts felt that the graph of the mass versus time for the electroplating experiment were both unrealistic and hard to read. Attempts were made to improve the representation of data, but student responses indicated there were still problems. The lines were too thick making it difficult for students to find a specific point on the graph. Units were added and the rate of silver deposition was modified to make it more reasonable. On the post assessment, experts felt that

students needed more information about how batteries worked. The scenario was changed per expert panel suggestions to relate the battery scenario to something that students understood. The scenario was expanded to both include an overview of how a battery worked as well as explaining how the terms in the graph related to the cycles.

The "Think-aloud" assessment required the most work. The word choice was deemed to be more of an advanced college level than a high school level. The experts felt that there was not enough information for students to adequately understand the material. They wanted an image of the silver nanoparticles as well as a more complete explanation of the science. There was confusion about whether we were asking students to understand the instrumentation or the science. The scenario was reworded to focus on the creation of silver nanoparticles creating an increase in absorbance. Images were added to show the nanoparticles as well as more explanation of the scientific process.

The graph was deemed to be unclear. The axes were relabeled in an attempt to make the graph clearer and additional words were defined. Even after these changes were made, there was a concern that students would not be able to interpret the graph. An additional summary graph was added in a further attempt to clarify the data and the question was rewritten a second time with clearer language.

The first three assessments were deemed to have high CVI scores and therefore high content validity. The "Think-aloud" assessment only had a CVI score of 0.78. Work was done on the assessment, but it was not resubmitted to the expert panel.

Discussion

Formative assessment and data analysis skills

Blank assessment findings: When the pre-assessments were given, a surprising number of them were blank or filled out with "I don't know." Both classes had approximately 50% blank assessments (Class A had 54% and Class B had 46%). The number of blank assessments could have been indicative of tasks that are outside the zone of proximal development (ZPD) described by Vygotsky (2011). Vygotsky described the zone of proximal development as the region where a student can learn with guidance such as scaffolding or mentoring. This region is where students have the highest learning gains. Outside that region are things that the learner cannot do. During the pre-assessment the following comments were heard: "I don't know how to do any of this." "I can't do this." "I don't get it." "This doesn't make any sense." These comments combined with the high number of blank assessments indicate that some students felt that they could not answer the questions in the assessment.

Another interpretation could have been that students were not motivated to complete the assessment. There are many theories about what motivates students (Brophy, 2010). Behaviorist theories talk about using rewards and punishments to motivate. In school, that can translate to grades (reward) or referrals (punishment). Students knew this assessment would not affect their grade and they were not likely to suffer any negative consequences for failure to complete the assessment. This implies that behaviorist theories of motivation are not useful in understanding why

students did not complete assessments. In Maslow's theory of needs, students work to fulfill needs. They must satisfy needs lower on the hierarchy before they can work on higher needs. The assessment does not map well to any of Maslow's needs so need theory will not help understand the situation. Goal theory implies that students complete work to satisfy some goal, but completing the assessments could not be linked to any goals the students had. There is only one motivational theory that seems to relate to what was happening in this study, Self-Determination Theory (SDT). This leaves intrinsic motivation theories.

SDT (Ryan & Deci, 2000) states that individuals can either be externally regulated or internally motivated. The behaviorist ideas of "reward and punishment" fall into external regulation. Internal or what they call intrinsic motivation stem from three things: relatedness, competence, and autonomy. Relatedness is the sense of being loved or cared for. In other words, students feel a sense of relatedness if they believe a teacher really cares about them. Competence is defined by SDT as feeling effective. It is related to how well an individual can perform a task. Students will perform a task if they think they can do a good job. Autonomy is a sense of volition and psychological freedom. Students will experience autonomy if they feel they have a choice in what they are doing.

Did the students have a sense of relatedness, competence or autonomy? Students had little reason to feel a strong sense of relatedness. Although the researcher was involved in the class as a student teacher, she had not yet taken over the full duties of the classroom. Her role up to that point involved enforcing classroom rules more than instruction. On top of that, in the month before the

assessment was given, she was unable to attend the class at all. Thus, students may not have been motivated by any personal connection they had with her. The student comments listed above indicate that students did not feel they could accomplish the task. This would indicate that they had low feelings of competence. They were not given any choices in the assessment so there was also little autonomy as well.

Pink (2009) expanded the original SDT idea of competence to mastery in his book *Drive*. He indicates that individuals will do a task they cannot do effectively if they feel like it is something that matters. As an example of this, motivation can be strengthened when students obtain an explanation for uninteresting tasks (Jang, 2008). When students were given the pre-assessment, they were given little explanation. Since students had little relatedness, competence or autonomy and had no explanation to motivate them, it is highly probable they were unmotivated.

There was some evidence during the cognitive interview to support a lack of internal motivation. During the cognitive interview, one of the students indicated that she did not know how to interpret the graph. She struggled with understanding the scenario and could not state in her own words what it meant. She asked if she could quit the assessment, but the researcher prompted her to continue. In the end, she was capable of completing three of the four problems on the assessment. This indicates that she had the ability to complete the assessment, but did not feel motivated to complete it. It is unknown how many students failed to complete the assessment due to their inability to complete the work and how many gave up on the assessment because they felt like they were not competent.

One way to differentiate between these two situations could have been to take class time to discuss it. If this had been done, students could have explained why they did not complete the assessment and those issues could have been resolved. The students could have retaken the assessment and allowed the research to go forward as originally planned. The researcher would also have known if blanks should be interpreted as lack of understanding or lack of motivation. Due to the short duration of the study, this discussion did not happen.

Instead, the curriculum was modified to provide additional instruction and scaffolding for the data analysis skills targeted in this study. Both classes received this instruction prior to the intermediate assessment, but a decrease in blank assessments was only seen in Class A, which received evaluative feedback prior to the intermediate assessment. Class A had only 29% blank intermediate assessments in contrast to the 43% for class B, which did not get feedback before this assessment. After both classes received feedback, the percentage of blank assessments was more closely matched. Class A had approximately 0% blank final assessments and class B had only 6% blank final assessments. This indicates that something definitely changed with respect to either understanding or motivation.

During the intermediate and final assessment, the following types of comments were heard: “Claim, evidence and reasoning? I got this,” and “What was reasoning again? Oh yeah, connecting evidence to the claim.” Even with these positive comments, a number of completed assessments still had incorrect answers. This suggests that the intervention did not impact student ability as much as student perception of their ability. Hume and Coll (2009) found that student self-esteem and

confidence to succeed is improved by the evaluative feedback that is part of formative assessment. Shute (2008) found that goal oriented formative assessment can have a large impact on motivation and student perceptions about how attainable goals are. Ryan and Deci (2000) state one part of intrinsic motivation is competence. Evaluative feedback helps students see what needs to be done in order to improve their work. This improves student's ability to do the work effectively, which would improve feelings of competence and therefore motivation.

The reduction of blank assessments seems to be more correlated to intrinsic motivation than actual data analysis skills development. The conditions at the beginning of this study seemed to lead to low feelings of competence and relatedness, which would have negatively impacted intrinsic motivation. During the course of the study, students received evaluative feedback, which improves feelings of competency. The increased feelings of competency would lead to higher internal motivation. This higher intrinsic motivation would lead to more work on the assessments. The reduction in blank assessments was greater than the learning gains that were measured through the course of the assessments. This implies that motivation was improved more than data analysis skills. However, since this study was not designed to investigate motivation, these findings are preliminary at best.

Intermediate assessment findings: The first question being addressed in this study was to determine if formative assessment could improve data analysis skills. There is some evidence that the feedback associated with formative assessment did in fact improve some of the data analysis skills targeted by this study. Class A received the treatment first and had statistically higher scores for graph

interpretation (scientific explanation) and hypothesis testing on the intermediate assessment.

There were no statistically significant gains in graph comprehension, inferences and patterns or error analysis seen. Graph comprehension had relatively high scores in both classes for assessments that were filled out. Students seemed to understand fairly well how to find a point on the graph. Observations in class indicate that even when students missed this question, it was often due to issues with graph clarity rather than inability to read a graph. Therefore, there was little room for growth in this area.

The area of inferences and patterns was affected by the grading rubric and lack of feedback. The rubric contained only three levels, 0-2, for each question. Leighton (2011) indicates that there should be four to five levels in a rubric if IRR is to be obtained. During the IRR sessions, there were many issues raised about grading the section on inferences and patterns with the rubric as it was given. When assessments were scored, IRR scores were below the 0.7 threshold value for this question. This IRR score, in addition to the questions raised during the IRR session, indicate that the rubric was insufficient for grading this group of questions. As a result, any gains seen would need to be reviewed and potentially rescored with a revised rubric.

There were no statistically significant gains seen for inferences and patterns. When the research design was altered, feedback was given on the science inquiry lab report. This report did not have a section that directly correlated to the inferences and patterns portion on the FA instrument. It would have been difficult

for students to transfer learning from the evaluative feedback on the report to the assessment instrument. Figure 2, which shows the frequency of comments given on science inquiry papers, has no comments related inferences and patterns. Since students did not receive evaluative feedback on these data analysis skills, it is not surprising that no gains were seen.

The lack of improvement on error analysis comes from a variety of sources. The first problem was the grading rubric. Although it had five levels, the levels didn't seem to match the question very well. The question was asking for data to "determine if the measurements were accurate", but this only really points to one type of error, measurement error. The grading rubric was looking for all four kinds of error listed by Hug and McNeill (2008): design, representation, execution and measurement. Since students were not prompted to provide multiple types of error, the highest levels of the rubric may have been unattainable. Indeed, most students scored zero or one, which points to an issue with either the instruction or the rubric.

The second problem with the error analysis question was the assessment itself. The "Think-aloud" interview showed that students interpreted this question incorrectly. Some thought the question was redundant and required the same answer as the hypothesis testing question (which was the question right before the error analysis question). The answers that were recorded on the assessments indicated that many students in class fell into this category. Some students provided the same answer for both questions and others specifically referenced the hypothesis-testing question in their answer to the error analysis question. Others

interviewees didn't know how to answer the question. On the final assessment, this question was the one most frequently left blank.

Another problem with error analysis was lack of scaffolding. Students didn't really understand how to determine limitations of data. They were given model papers but these models did not clearly outline the different types. Different sources of error were discussed in class, but the final revisions of the science inquiry lab reports indicate a lack of understanding around where error could come from.

On top of this, students were dealing with second hand data on the assessment instruments and first hand data in their science inquiry lab reports. Hug and McNeill (2008) indicate that students talk about execution and measurement limitations when discussing their own data (first had data) but do not normally talk about sources of error with second hand data. In the initial plan for this study, students would have gotten evaluative feedback on the assessments instruments. This would have helped students understand how to look for limitations in second hand data. When the study was changed, students received evaluative feedback on their science inquiry reports, which only included first hand data. Since students don't normally consider sources of error with respect to second hand data, it is unrealistic to expect them to improve data analysis skills on second hand data without some intentional instruction. This instruction was not provided so it is unrealistic to expect improvements on error analysis.

In summary, statistical learning gains were seen for graph interpretation (scientific explanation) and hypothesis testing. Changes to the study, lack of appropriate scaffolding, and inadequacies in the grading rubric contributed to the

lack of gains in the areas of inferences/patterns and error analysis. Additional research would need to be done to see if FA can aid in learning all these skills.

Final assessment findings: At the end of the study, both classes were given a final assessment. This final assessment was intended to show that both classes were matched at the end of the study and that all differences on the intermediate assessment were due to the evaluative feedback. This was not the case. Class B, which received delayed treatment did not achieve the same gains. There are multiple potential reasons for this.

The first issue had to do with the formative assessment process. The initial design of our experiment involved providing feedback of student performance on the assessment instruments. Because there were such a high number of blank pre-assessments, it was not possible to give meaningful feedback to students using the instrument. Continuing the research without modification would have changed this research to be about a tool instead of the full FA process. Multiple sources indicate that FA as a tool is not as effective as the full FA process (Hume & Coll, 2009; Shute, 2008; Karee & Sean, 2009; Dunn, & Mulvenon, 2009). To maintain integrity in the FA process, something needed to change.

The decision was made to modify the experimental design and give feedback on the science inquiry lab. When this change was made, the number of opportunities for feedback was reduced. Students, who received the delayed treatment, obtained indications about proficiency before they were given feedback instead of receiving one assessment with scores and one assessment with feedback. Unfortunately, the

literature indicates students do not internalize feedback as much when they receive scores first (Hume & Coll, 2009).

The literature also indicates that feedback on skills needs to be more immediate than feedback on content (Shute, 2008). In the original design of the experiment, students were supposed to get immediate feedback on the assessments. Class A was supposed to receive feedback on the pre-assessment. Class B, which had a delay in treatment, would have gotten immediate feedback on the intermediate assessment. When the research was changed, the FA process changed to include the whole instructional unit instead of being confined to the instruments. This resulted in fewer feedback opportunities. Instead of giving Class B immediate feedback on a later assessment, they received a delay in feedback on the same assessment, the science inquiry lab report. This delay may also have impacted the ability of students to incorporate and internalize the feedback. If students did not internalize the feedback, they may not have gained many of the benefits. These two issues, providing indications of proficiency before feedback and a delay in feedback, may have interfered with Class B's ability to benefit from feedback, which would have resulted in lower scores on the final assessment.

Another reason Class B may have scored lower on the final assessment than Class A may have been demographics. While both classes were mixed ethnicity, the set of complete assessments (students that had completed all three assessments) was filled out by a less diverse group of students in Class A than Class B. In Class A, 67% of the completed assessment sets were done by students whose first language was English. In Class B, only 43% of the complete assessment sets were done by

students whose first language was English. Since the assessments were heavily text driven and required a lot of reading, this may have hampered student's ability to understand both what was happening and demonstrate data analysis skills learning. According to Bloom's revised taxonomy (Anderson et al., 2001), students need to be able to understand material before they can analyze it.

This last issue with demographics suggests that the assessments may have been biased towards native English speakers. To resolve this, assessment instruments need to be modified to make them more accessible. Of the multiple techniques listed in Janzen's review (2008) of English Language Learner literature, only creating comprehensible input and multiple forms of input are relevant to these assessments. Before any future work is done on these assessments, language should be clarified and additional modes of input should be added such as reading scenarios out loud and adding more graphics.

In addition, data analysis skill instruction needs to be more accessible as well. The instruction method for scientific explanation was also very text heavy. It would benefit all students if instruction around scientific explanation involved more than text analysis and one practice session. If expectations for scientific explanation are built into the classroom norms, students will construct their own definitions for terms like "claim", "evidence" and "reasoning," which is considered best practice for teaching English languages learners (Fisher, Frey, & Williams, 2002; Janzen, 2008). In addition, Driver et al. (2000) indicates that scientific explanation needs to be learned in a social situation, which means this practice would help all students.

In summary, the preliminary results indicate that the evaluative feedback in the formative assessment process does help students gain some data analysis skills. In addition, there are indications that the feedback may also have helped either student motivation or competence. There are a number of limitations to this research that were caused by changes to the experimental design, a demographic mismatch between classes and inadequacies of the grading rubric. Thus additional research is needed to confirm these findings.

Student thinking when approaching new tasks

There were two sources of information used to evaluate student behaviors and thinking when addressing new tasks. The first was the science inquiry lab shown in Appendix B and the second was the cognitive interview that was done with the "Think-aloud" assessment from Appendix D.

Science inquiry lab: The science inquiry lab exposed three types of behaviors: engagement, questioning and avoidance. Students familiar with the task became engaged quickly only asking for locations of certain chemicals. A portion of the students who were unfamiliar with the task asked questions until they were able to accomplish it. Their questions were often around defining what they were being asked to do and helping them discover how to do it. The last group avoided the task until an instructor gave them explicit step-by-step instructions on how to complete the task.

What made this last group different? Many students in this last group coincided with what would be considered low performing students. When

confronted, these students indicated that they did not understand the task, but they made no effort to understand the task either. This suggests disengagement and a lack of motivation. This lack of motivation can be directly tied to what the students said. People need to feel a degree of competence in order to be self-motivated (Ryan & Deci, 2000) and these students did not feel they could effectively complete the lab with the amount of instruction they were given.

To understand this, it is important to discuss the norms of the classroom before the study was conducted. Discussions with the cooperating teacher indicated that students had been doing inquiry, but there was a miscommunication between the researcher and the cooperating teacher about the definition of inquiry. Prior to this guided inquiry experience, the students in this class were predominantly exposed to expository labs, where everything was directed, or structured inquiry, where the questions and procedures were directed and only the outcome was unknown. Unfortunately, Hume and Coll (2009) indicate these kinds of step-by-step instructions only help students learn science process skills instead of higher order critical thinking skills. The researcher was interested in doing guided inquiry or open-ended inquiry in order to access higher order thinking skills.

There are many proponents of open-ended science inquiry, but most agree that science inquiry is a process that must be learned. (Bell et al., 2005; Berg et al., 2003; Hofstein & Lunetta, 2004). Students can't be expected to jump into open-ended science inquiry without preparation. In the case of this class, some students found it even too big of a jump to go from the structured inquiry they had experienced to the guided type inquiry that was being done. The disengagement of

these students is explained by SDT (Ryan & Deci, 2000). Students had greater autonomy, they were allowed to choose which reaction they were doing and what factors they were testing, but they didn't feel competent. Students need to have both autonomy and competence in order to be intrinsically motivated. Since competence was missing, students were disengaged from the process.

Wheeler and Bell (2012) suggest a number of questions to help students through this transition including a gradual removal of scaffolding. Because of the communication between the cooperating teacher and the researcher, there was not enough scaffolding to help all students make the transition from the more directed structured inquiry to the less directed guided inquiry. The end result was that these students were dependent on the teacher and the task was reduced to a structured inquiry lab. When the task was lowered to a more procedural level, students expressed a higher level of comfort with the task. They also did not use as much higher order thinking.

There are multiple ways that this issue could have been avoided. Additional scaffolding could have been provided. Students could have been given exercises to help practice and discuss how to plan and create procedures to test a hypothesis. For example, they could have had to order jumbled lab procedures or been asked to write procedures from something that is familiar to them. More in class planning time could have helped. If students were given more time, the researcher or cooperating teacher could have walked through student science inquiry plans to ensure they were viable. Lack of student understanding would have been caught and additional instruction could have been provided. More practice with inquiry

could also have helped. Studies show that student performance on science inquiry increase with practice (Marx et al., 2004 & Berg et al., 2003). In other words, future work should involve more science inquiry experiences that transition from structured to open-ended inquiry through the slow removal of scaffolding.

Science inquiry lab report: Writing the science inquiry lab report was outside what was normally expected for this class. The cooperating teacher had assigned what he called literacy pieces, but these predominantly were one page assignments where students were expected to write two sentences that fit a given pattern on a particular topic. For example, one assignment asked students to list two substances and indicate which was more acidic. Students were then asked to indicate two tests that could be used to determine which of the two substances was more acidic. In the instructions given, it said to find the answers in their notebook and copy them into sentence frames. All of this information could be obtained from one lab.

This example shows that higher order thinking was not one of the classroom norms. According to Blooms revised taxonomy (Anderson et al., 2001), asking students to review their work, evaluate the results, and explain the findings should have required higher order thinking. The verbal instruction given in class and written instructions on the paper indicated that students should just copy answers in their notebooks instead of performing any of these higher order thinking tasks. These additional instructions lowered the cognitive demand of the task to remembering, which is the lowest form of thinking on the revised taxonomy. Students only had to remember which lab related to acidity then read the note-taking guide they were given to find the answers. Lovett and Chang (2007) found

that students avoid higher order thinking if they can find the answers using queues from the questions themselves. The instructions given by the cooperating teacher provided enough clues that students did not have to do any work. This resulted in a situation where the cooperating teacher felt like higher order thinking was one of the norms of the class, but in practice it was not. This disconnect was evident in the way students approached many of the tasks that were part of this study including the science inquiry lab report.

Students were given instruction on writing a scientific explanation in class. They were provided with models that showed what a good report looked like and given feedback on how to be proficient. In the end, only eight of the final papers were proficient. Of the original 45 not proficient papers that were turned in, only 26 were corrected and turned back in. One additional paper was completely rewritten without incorporating any of the feedback and then turned in. This shows a high degree of avoidance on the science inquiry lab report.

Why did this happen? First, as was stated above, higher order thinking was outside the norm of the class. In effect, this assignment was asking students to do something they were not accustomed to doing. Second, the feedback itself should be examined. The types of evaluative comments given on the science inquiry lab reports were predominantly negative. They indicated what was missing. Some contained comments indicating that some part was good, but these positive comments were more likely to be on papers that were closer to proficiency. Looking at Figure 2, it is clear that the students that chose not to complete a final draft received a higher volume of negative comments than the students who chose to

complete a final draft. Self Determination Theory (SDT) indicates that negative comments can negatively impact motivation (Ryan & Deci, 2000). It is possible that the students who did not complete the assignment felt like they were less competent because of the comments. If students do not have any intrinsic motivation, they only complete projects based on external forces. As was stated before, many of the students who did not turn in final drafts were low achieving students. These students were struggling to pass the class. They needed to decide where to invest their limited time and energy if they wanted to pass the class. If this assignment could not make them pass, there was no reason to complete the assignment. Finishing a final draft was not worth the effort.

Another factor that impacts intrinsic motivation is relatedness (Ryan & Deci, 2000). The researcher explained feedback, but she left the next day to take another job. This may have impacted any feelings of relatedness students had with the researcher. It is hard to maintain a connection with students without remaining in their classroom. In addition, any follow-up questions would have gone through the cooperating teacher, who may have approached the questions from a different perspective. This may have confused students causing them to feel less competent. The cooperating teacher may have had a different idea of what proficiency looked like. He may have guided them to create non-proficient papers that only incorporated some of the feedback.

The lack of student participation on the science inquiry lab report (only 50% final drafts were turned in) raises the question, “What does it take to establish a classroom where students incorporate feedback and improve?” In this study,

students were told that the final science inquiry report would impact their grade. According to SDT (Ryan & Deci, 2000) this would be external coercion. This type of pressure would only obtain compliance from students who already were motivated to care about their grade. Gaining participation from a fraction of the class falls short of the goals that are set for education (NRC, 2011). To gain participation from all students, motivation needed to become intrinsic. Ryan and Deci (2000) indicate there are three pieces of intrinsic motivation: autonomy, competence, and relatedness. Students could gain some autonomy. Autonomy can be summarized by the following questions: what, when, where, and how. Giving students autonomy would mean they have some choice over what they work on, where they work on it, when they work on it and how they do it. The way that school is structured at this time gives little room for changing the timing of when students can work on things, but there is room for students to decide what they want to work on within boundaries, how they work on accomplishing goals and where they work on projects within the room. In a classroom, there could be multiple modes of presentation that students could pick. The teacher would just have to ensure all the form of evaluation were equivalent. In the case of open-ended science inquiry they are given options about what they study as well. Students need to learn certain concepts, but there can be flexibility on how they learn those concepts. It means the teacher has to give up a little of the control in the room.

Competence is a measure of how well individuals feel they can do at an activity. In many ways, scaffolding helps build a sense of competence. If class work is scaffolded well, students understand what the goal is and understand how to get

there. Evaluative feedback seems like it could help students build confidence as well, but it has to be done right. Purely negative comments reduce feelings of competence. If students only hear what they have done wrong and not how to apply it to get better, it can reduce feelings of competence (Ryan & Deci, 2000). One idea would be to create mix of positive and negative comments in the evaluative feedback so that students understood how to raise their performance while still making them feel capable of raising their performance. The problem is that students may not listen to the good feedback when it is paired with negative feedback. In fact, if students get a bad grade, they often do not internalize feedback (Hume & Coll, 2009). Teachers need to find a way to make feedback constructive instead of simply critical.

One way to do that is through relationships. The last factor that influences intrinsic motivation is relatedness. Relatedness is both a feeling that someone cares and a sense of shared experience. The FA literature actually hints at this idea. Black and Wiliam (1998) talk about teachers partnering with students towards success. This could make the experience in the classroom more of a shared experience instead of something students have to live through. Partnering with students would also help students feel like teachers care.

Building relationships fosters intrinsic motivation, but how is this done? Why do some students take feedback as criticism and other interpret feedback as help? For students to partner with teachers towards learning, the classroom must be established as a learning community. Brophy (2010) gives three important ways to accomplish this task. The classroom must be an attractive place. The teacher needs to be welcoming, cheerful, sincere and interested in the students. This goes right

back to the definition of relatedness. Students must feel like teachers care. Student's attention must remain focused on collaborative learning goals. This means teaching concepts that are worth learning that lead towards concrete goals and powerful ideas. Instruction must be altered in response to student learning. Finally, lessons need to be taught with a goal towards understanding. Unfortunately, this study was only designed to last one month and one month is not enough time change classroom norms and create a learning community. This type of learning environment needs to start at the beginning of the year and exist all year long.

There are short-term fixes that could be used. Jang (2008) indicates that providing explanation can increase motivation. In his study, the explanation provided explained how the task would help students master the concept of analyzing data. It goes onto tell them why mastering the skill is important, they will be better teachers. This fits in with Pink's (2009) extended version of competence, what he calls mastery. People want to feel like they can get better at something that matters. He also extends the idea of relatedness into what he calls purpose. He defines purpose as being in service to something larger than themselves. This suggests that students would be more likely to incorporate feedback if they felt that the final project helped other people and made the students more connected to the world at large. This fix will only work long-term if a learning community is established. Otherwise explanations to motivate students become insincere.

In summary, there was a lack of student participation. Half of the class did not turn in final drafts of their science inquiry lab reports. This lack of participation suggested that students were not intrinsically motivated to complete the task. In

order to improve intrinsic motivation the classroom culture needs to support a learning community. Creating a learning community cannot happen in one month. It needs to start at the beginning of the year and be supported all year long.

"Think-aloud" interview: The "Think-aloud" interview was done with six different students that were supposed to represent three different ability levels in the class: high performers, average performers and low performers. Recall that these designations were based on how students had performed in the class prior to the study as well as how students performed on the assessments. In these interviews, students were asked to complete an assessment while talking about their thinking processes. Students from different ability levels approached the assessment in different ways.

High performing students: The high performers carefully read through the scenario and continuously referred back to the scenario in order to answer the questions on the assessment. While the high performing students had high average scores on the final assessment, their performance dropped on the "Think-aloud" assessment. Both students commented that they didn't have enough information in the assessment to understand what was happening or answer the questions. In support of the students, it should be noted that this was the only assessment that did not reach the 0.8 CVI threshold for content validity specified by Rubio's (2003) paper.

It should also be noted that the "Think-aloud" assessment had the least in common with student experiences. The scenario was unfamiliar and was not closely related to the curriculum content. In Bloom's revised taxonomy (Anderson et al.,

2001) understanding is below analyzing, which implies that students cannot perform higher order analysis on information they do not understand. On previous assessments, there was evidence that students were using content material from the class to augment the scenarios given in the assessments and provide additional information. This suggests that testing of higher order data analysis skills cannot be done without considering content knowledge. Further research is needed to confirm this statement.

Both high performing students that were interviewed expressed a desire to have the raw data present so they could manipulate the data themselves. One student even indicated that he understood ideas better if he had a chance to work with the data instead of just a representation. While it was not feasible to provide all the raw data, perhaps a table that summarized the data instead of an additional graph could have provided more meaning to the higher achieving students. It also may have helped to provide more physical models of the scenario instead of just a diagram so students could understand what was happening better.

It was also notable that the high performers seemed to have more persistence than the other groups. They spent almost twice as much time on the assessment as the low performers and interacted with the text more in order to gain meaning (underlining terms and rereading sections). According to SDT (Ryan & Niemiec, 2009), a high level of persistence is indicative of higher internal motivation. This also suggests that the other groups, average and low performers, had insufficient motivation to complete the task at the same level. As was stated earlier,

this points to a deficit in of competence, relatedness or autonomy in the average and low achieving students.

One of the two high performing students said she would have to do the experiment herself to answer the question about error analysis. This points to issues with first-hand and second-hand data. Each of these assessments relied heavily on second-hand data, which is common for assessments that are not imbedded in the curriculum. Instead of collecting data themselves (first-hand data), students read about scenarios and data collection that was done by someone else (second-hand data). Hug and McNeill (2008) indicate that students are less likely to discuss data limitations or sources of error on secondary data.

Students need to be able to find limitations in secondhand data. Standardized tests use secondhand data and most of the information students read on the internet or hear on TV is secondhand data. Without being able to assess the quality of secondhand data, it is difficult to determine the validity of scientific explanations. How can a student judge the quality of a scientific explanation if that student cannot judge the validity of the data supporting the claim? Indeed, the NRC (2011) indicates that students need to be able to make judgments about conflicting scientific claims. How do we get to a point where students can do this?

Hug and McNeill (2008) suggest using inquiry projects that involve a large degree of secondhand data. Hug and McNeil give the example of evolution, where it is unreasonable to generate data sets in the classroom. There is little research on how to improve student evaluation of secondhand data. One potential for future research would be to create scaffolded evaluation of secondhand data. For example,

students could analyze extensions to their data. Students could do a lab, then be given additional data collected by someone else and asked to determine the quality of the data based on their own experiments. This last suggestion would need addition study.

Average performing students: Both of the average performing students tried to initially answer questions without reading the scenario. They focused on the graph instead. One of the average performers went back to read the scenario once she began struggling with the questions. She did not read it completely at first. She skimmed through the words looking for something that she could write down. Later she realized her answer was wrong. At that time, she fully read the scenario. The other average student never read the scenario. He used only evidence from the graph such as "one line is going up and the other going" down as their claim. His evidence and reasoning were "the graph shows it." He didn't try to understand what the graph meant, what the numbers on the graph represented and didn't bother to even read the title of the graph. When asked about it, he said he always tried to answer questions based on the graph. He thought the scenario contained "too many words" and didn't want to go through it.

Both of these students demonstrate behavior similar to Lovett and Chang's (2007) findings. The students were using queues from the questions and images to find answers instead of trying to comprehend the overall scenario. This behavior may have been supported the question wording itself. For example, the second set of questions labeled "Graph Interpretation" asked students to explain what was happening in the graph using their own words. Students may have interpreted this

question to mean they didn't have to read the scenario. Since this was not the intention of the question, the graph alone did not contain enough information to create a meaningful scientific explanation. Explanations created using the graph alone did not score highly on the rubric.

This behavior, attempting to answer questions without reading, may also have related to the classroom norms. The class was structured around activities with very little reading. Students did not use a textbook and the cooperating teacher ensured that both the reading level and the amount of reading on handouts were minimized. Students used their lab notebooks to record observations and take lab notes, but they were not interactive tools. Notebooks were simply places to store facts for easy retrieval later. In this environment, it is possible that students felt like they could learn and do well in science without reading. Students may have realized that they needed to read the scenario if they had been given feedback on the assessments. However, it is not clear that simply providing evaluative feedback on previous assessment instruments or changing the wording of the question would have helped the average performing students on the "Think-aloud" assessment instrument.

The average performing student who did read the scenario, didn't understand most of it. She consulted the definitions at the bottom of the assessment, but did not seem to comprehend what they meant. She read through the scenario several times, tried to put the defined terms into sentences from the scenario and shook her head. Later she commented that the definitions didn't really help. She understood the words that were part of the definition, but couldn't put the words

together into a meaningful understanding of the term being defined. It should be also be noted that both of the low performing students struggled with comprehension issues as well.

The behavior of the average and low performing students highlights an issue with the assessments. The assessments were created using the assumption that more difficult terms could be defined, but this many be insufficient to support higher order thinking. The book *Reading and Writing in the Content Area: Practical Strategies* (2007) states that students do not always understand definitions of vocabulary words if they are just presented as text. Vocabulary must be taught. The book goes on to provide a number of methods to teach vocabulary including concept definition maps, discussion around difficult terms, and structured overviews. All of these involve more interaction than students had on the data analysis FA instrument. Since it is unfeasible to teach vocabulary during an assessment, assessments could be rewritten at a lower reading level so there are no terms that need to be defined. A better option would be to embedded FA into the curriculum or aligned FA to the curriculum so that the terms are already explored and defined in class.

This problem, difficulty comprehending defined terms on an assessment instrument, is compounded when dealing with English language learners. Fisher et al. (2002) indicates that teachers must do more than provide definitions. Teachers must help students develop transportable vocabulary skills by reviewing the definitions of the words in different contexts to expand and deepen the meaning. Barr et al. (2012) indicates that words must be placed in context so that students can use contextual clues to understand the words. In other words, students used the

scenarios to help them understand the definitions rather than using the definitions to help understand the scenario. If the scenarios are left with terms defined at the bottom, there is a risk that they will be biased against English language learners. The assessments then become about reading comprehension instead of data analysis skills. This issue may not have been apparent to the expert panel when they considered content validity, because few of the panelists had direct experience with English language learners. Those panelists that did have experience with English langage learners may not have realized what portion of the students in this study did not have English as their first language.

Low performing students: Both low performing students initially attempted to answer the questions on the assessment without reading the scenario. Quickly it was apparent that they could not succeed. Both students read the scenario, but there was little evidence of comprehension. After reading through the scenarios, both students tried to select certain terms and randomly fit them into their answer in ways that did not make sense. In one case, the student confused the independent and dependent variable. In the other case, the student used words in ways that didn't make sense.

When asked about their behavior, one student commented that he didn't know how to read science. As was stated earlier, the class was structured to minimize reading. Unfortunately, the book *Reading and Writing in the Content Area: Practical Strategies* (2007) indicates that reading in any content area is a skill that must be developed through practice. By altering the class to minimize the reading requirement, the cooperating teacher was also limiting student's ability to learn a

crucial skill. The NRC (2011) indicates that students need to be able to evaluate and construct arguments. If students are not exposed to written science, their ability to evaluate it is limited. Additionally, if the assessments are a higher reading level than the students understand, the assessments can become more of a reading comprehension assessment than a data analysis skills assessment.

One of the lower performing students rushed through the assessment and made multiple mistakes as a result. He kept going back to change what he was doing and actually completed the assessment twice in less time than it took a high performing student to complete it once. He made several comments ("Oh I get it.") as though he comprehended something but his answers ended up more confused than when he started. There was one moment when he stated several valid ways to answer the question about testing a hypothesis, but wrote down an answer that was incomprehensible. His answers suggested that he didn't really know how to perform analysis suggesting that these skills should be directly addressed.

Trends across all ability groups: Students in all ability groups demonstrated that they were learning from the assessment. Several students read questions that caused them to go back and change earlier answers. In one case, the student commented that his reasoning didn't make sense. He requested a new sheet so he could start taking the test over from scratch so he could change his claim, evidence and reasoning. In another case, the student read a word on the second to last question that she had previously not understood, nanoparticles. She went back and read the scenario placing more emphasis on the word nanoparticle and adjusted her scientific explanation to include this idea. A third student commented that he felt

like he learned more about the topic as he was completing the assessment. He felt that he would do better if he were given the assessment again.

Crooks (1998) noted that tests could impact the way students study and learn. Many students used tests to determine the real goals for the class. In addition, Crooks reported that students who had frequent testing had higher performance. This idea that students learn from assessments is one of the key aspects to FA (Black & Wiliam, 1998). In FA, the students are being guided to learn from the assessments through the evaluative feedback. During the "Think-aloud" interview, students received no guidance. Learning on the "Think-aloud" interview was driven mostly by the students themselves. In this case, high performing students learned most effectively than low performing students. Why was this the case? According to SDT (Ryan & Deci, 2000), higher performing students usually have a high degree of internal motivation. This internal motivation could lead to what Pink (2009) calls a need for mastery. Mastery is the process of working to get better at something that matters. Comments made by high performing students during the "Think-aloud" interview indicated that understanding matters to these students and higher order thinking leads to a higher degree of understanding. Thus, these high performing students may have more practice in the higher order thinking involved in these tasks than their lower performing counterparts.

The "Think-aloud" interviews demonstrated multiple issues with the assessment. The assessment did not have enough information, had questions that prompted students to answer questions using faulty methods, and was written at too high of a reading level. In spite of these setbacks in this research project, some

information was gained. Successful students were more likely to utilize literacy strategies such as rereading text, underlining passages and trying to understand meaning from contexts. Average and less successful students attempted to complete the assessment without understanding the scenario. Students also struggled with the vocabulary used in the assessment. These findings suggest that FA in conjunction with literacy strategies could improve data analysis skills more than just FA alone. Additionally, this study highlighted issues with using secondhand data to assess student data analysis skills. Curriculum should be developed to help students learn how to properly evaluate second hand data in order to meet the goals outlined by the NRC (2001).

Validity and reliability of assessments

The final question that was investigated by this study involved the validity of the assessments used. All assessments were reviewed by an expert panel and most received an overall CVI score of >0.8 (Rubio, 2003), which is considered to be above the threshold for content validity. The "Think-aloud" assessment did not reach this threshold. During the "Think-aloud" interview, students determined that the "Think-aloud" assessment lacked the information needed to answer the questions. When the other assessments were reviewed, it was determined that some of them lacked information needed as well. Why was this the case?

One possibility is that the rubric to score the assessments was created after the assessment instruments. According to the best practices of backwards design

(McTighe & Wiggins, 2005) goals (such as those in a rubric) must be solidified before assessments and instruction is created. Since this was not done, the expert panel had no way of aligning on what proficiency meant. The expert panel may not have been able to ensure the quality of the questions, which would in turn make the content validity scores less accurate.

More likely, is that content validity alone is not enough. Recall that Cronbach, & Meehl, (1955) define content validity as showing that the questions are a sample of the universe that is being investigated. For this research, the CVI (content validity index) score was a measure of how well the questions tested data analysis. While this was an important piece of information, it was not the only important piece. In fact Messick (1990) indicates that multiple measures of validity are needed to ensure assessments are appropriate.

Student responses on the intermediate, final and "Think-aloud" assessment support the claim that content validity was not enough. Student responses indicated that multiple students misinterpreted questions. Students interpreted the question about error analysis on the scenario to be about hypothesis testing. Students tried to create a scientific explanation about the graph instead of the scenario. Some students seemed to misinterpret the question about hypothesis testing as well. When subjects interpret a question to mean something different than the researcher intended, Willis (2005) calls this response error. In his book, he suggests performing cognitive interview before using instruments to improve clarity of questions and reduce response errors. When the study was designed, the interviews were placed at the end as a confirmation of content validity. The high volume of response error

suggests the cognitive interviews should have been done in the beginning to ensure clarity of the assessment.

Why didn't the expert panel recognize there would be such high response error? They were asked about the clarity of the assessment. They were also asked whether questions, like the error analysis question, could adequately test student's ability to determine sources of error. It is possible that the experts could not view the assessments through the same lens as the students taking the assessment. The "Think-aloud" interview indicated that the expert panel suggestions actually increased response error. The panel was comprised of PhD chemists, professors of education and highly experienced teachers. None of them have the perspective of a teenage student. Additionally, the expert panel reviewed the assessment in chunks through an on-line survey. It is possible that format hid some of the confusion that ended up being part of the assessment.

It is also possible that the expert panel needed to know more about the context to be able to truly evaluate content validity. Ruiz-Primo, Li, Tsai, & Schneider (2010) indicate that students rely on all of their knowledge to create scientific explanations. Without knowing the curriculum, it may have been difficult for the expert panel to know if students could adequately answer the questions. In fact, some comments from the expert panel indicated this was an issue. The expert panel was uncertain if the students had exposure to the terms "claim, evidence and reasoning." One panelist asked if students had covered equilibrium so they could apply that knowledge to the intermediate assessment and another wondered how much knowledge students would have about electrochemistry. Even if curriculum

materials could have been provided to the expert panel, they might not have given an accurate picture of the context. Initially, the assessments were designed to match with the curriculum, but changes to timing meant that they were no longer well synced. The only way to ensure that the expert panel had an accurate view of the context would have been to change the timing of the research. Unfortunately, this was not feasible.

Another way to ensure that students have the necessary background knowledge would be to create assessments that match more closely with the Next Generation Science Standards. The assessments that were created focused heavily on electrochemistry, which does not fit into the standards as well as reaction rates (standard HS-PS1-5.) or thermochemistry (standard HS-PS1-4.). Choosing more core topics would also mean spreading the research out over more than a month giving time to really explore multiple science inquiry labs, discussions around error analysis and practice on creating scientific explanations.

Improving the content validity, resolving issues with response error, and ensuring an appropriate context for the assessments would not have been enough to make the assessments valid and reliable. There was one additional issue that became apparent as the assessments instruments were scored. The assessments were designed to be parallel, but may not have correlated well with each other. Although the questions were the similar, the scenarios were different. Small differences in the scenarios could have led to different levels of student understanding. One piece of evidence demonstrates that the assessments were not

as well correlated as planned. The final assessment seemed to connect with the students more than the other assessments.

Hug and McNeill (2008) indicate that students use their experiences to help them understand scientific phenomenon. Berglund and Hammer (2012) indicate that student experience of science affects their ability to create arguments about it. The final assessment contained a comment that was different than the other assessments, "If you have an old phone, you might notice a giant drop in charge if it is not plugged in overnight." This comment directly related the assessment to the student's lives. When this was done, it highlighted a new data analysis issue that was not seen on previous assessment, use of everyday analogies as reasoning instead of science (Hug & McNeil, 2008). By connecting the data to student personal experience, it opened the door to using that experience in the explanation. Students then used inappropriate data to support conclusions (Ruiz-Primo et al., 2010). Student responses often included statements like "because that is what my phone does." Since the students lacked connection on the other assessments, they could not demonstrate this common error in data analysis. Feedback would not have improved understanding because they would never have gotten feedback on this point.

In summary, this study looked at content validity, but there is evidence that content validity was not enough. Additionally, the content validity may not have been as good as the CVI scores indicated because the experts did not have the grading rubric, did not know the educational context and did not necessarily view the assessment instruments as a whole. There was high response error due to

misinterpretation of questions. “Think-aloud” interviews were done to confirm findings and it might have been more effective to use them to improve the assessments. The assessments did not all correlate to each other. Future research will need to involve revised assessments. In addition, the revised assessments will need to have multiple kinds of validity checked.

Limitations

Measurement issues: IRR scores only reached the threshold of 0.75 for 3 of the five question groups. While attempting to obtain IRR, it was clear that the scoring rubric needed work. Most questions were graded on a scale of 0-2, but raters could differentiate more levels than that. For example, there was concern about scoring a student who had small misconceptions in the same grouping as students who had major misconceptions. This made scoring these questions problematic. Leighton (2011) indicates that rubrics need four to five levels in order to obtain reliable scores. Although gains were seen in the categories of inference/patterns and error analysis, these scores should be considered preliminary. The Rubric used for grading did not have enough levels to allow raters to differentiate levels of understanding.

The assessment instruments used in this study were designed to be parallel, but the evidence indicates that they were not. The final assessment connected more closely to the student’s lives than other assessments and the “Think-aloud” assessment did not reach the 0.8CVI threshold specified for content validity in

Rubio's (2003) paper. This mismatch may have affected comparisons made between scores on different assessments.

There was a high degree of response error, which is when subjects misinterpret assessment questions. When students answered the wrong question, their answers counted as zero on rubric. These scores did not reflect student ability to perform data. Since the questions that caused confusion were parallel for all assessment instruments, these errors ended up propagating through all assessments. Willis (2005) suggests performing cognitive interviews as a way of clarifying questions to reduce this error. Performing these interviews prior to the using the assessments could have lead to more accurate evaluation of student data analysis skills.

Design issues: The study was originally designed to provide feedback on the assessment instruments. When 50% of the pre-assessments were left blank, this original design could not be followed. The decision was made to change the FA process to include the science inquiry lab report. This limited the opportunities for feedback, introduced feedback delay for one class, and did not allow feedback in crucial areas like limitations of secondhand data. As a result, any findings must be considered preliminary and should be confirmed with a more rigorously designed study.

The study was planned for a short time period, less than one month. This did not allow for students to learn key skills like science inquiry or error analysis. A lesson was included on creating scientific explanations, but this could not be repeated in the timeframe. In addition, this did not provide for enough time to

create the kind of classroom environment where all students internalize feedback. Results suggest that one-month is not long enough to have a meaningful impact on data analysis skills or higher order thinking. Future research should span a full class year.

The study was only designed to evaluate content validity. Although three of the four assessments reached the 0.8 CVI threshold, there were still issues seen with the assessments. Most validity experts recommend obtaining multiple forms of validity (Cronbach & Meehl, 1955; Messick, 1995). Pursuing different kinds validity may have created better assessment instruments that were free from some of the measurement errors listed above. Additionally, assessment instruments were modified to incorporate expert panel feedback, but were not reviewed again by the expert panel. It is possible that modifications did not address issues that were brought up by the panel. This would have been highlighted if experts reviewed revisions.

Execution issues: Equal gains were not seen in both classes at the end of the study. One limitation that could have contributed was researcher participation. Feedback was given to Class A at the end of the researcher's direct involvement in the class. For the month leading up to the feedback, the researcher had taken the lead in all instruction for the period of the study. Immediately after feedback was given to Class A, the researcher was moved to a different school. Although the researcher came back to give feedback to Class B, the relationship between the researcher and the students may have weakened in that time. This may have caused

students to be less likely to either internalize feedback or fully understand feedback given by the researcher.

Two question groups (inferences & patterns and error analysis) did not have precise definitions in the scoring rubric. This suggests these data analysis skills were less defined in the researcher's mind. This lack of specificity may have propagated into the lessons themselves making the feedback for the data analysis skills less informative and the direct instruction less clear. To remedy this issue, it would have been better for the researcher to create the rubric before the assessments were given. Wiggins and McTighe (2005) indicate that rubrics need to be created before the task begins to ensure instruction and feedback was consistent with the goals. Once this initial rubric was created it could have been reviewed by the expert panel to ensure that definitions of these skills were precise and well understood.

The timing of the curriculum changed, but the assessment schedule did not. Students use information from all of their experiences to create scientific explanations (Ruiz- Primo, 2010; Hug & McNeil, 2008). When the timing of the curriculum changed, students no longer had a large amount of classroom learning to support their development of scientific explanations. Skills like scientific explanation and error analysis rely on a deeper understanding of the topic that can only be gained with sufficient content knowledge. Without this understanding, Bloom's revised taxonomy (Anderson et al., 2001) indicates students may not have been able to fully participate in the higher order thinking targeted by this study.

As a result of all these limitations, any findings in this study should be considered extremely preliminary and require additional research.

Future Research

The results of this research exposed many limitations including the short duration of the study and the validity of the assessments that were used. For future research, the assessments must be revised to address both of these issues. First the timeline must be adjusted. In this study, only one science inquiry experiment could be performed because the research only encompassed a few weeks. Inquiry is a skill that must be learned through multiple experiences that build on each other and move from structured inquiry to open-ended inquiry (Berg et al., 2003; Hofstein, 2004; Ruiz-Primo & Shavelson, 1996; Keys et al., 1999; Wheeler & Bell, 2012). In order to allow for multiple inquiry experiences with increasing levels of inquiry, the research needs to span a full term or a full year.

If the timeline is expanded, the scenarios within the assessments also need to be adjusted. The current assessments focus on topics that are covered near the end of the academic year. Bloom's revised taxonomy (Anderson et al., 2001) indicates that students must understand the meaning of written passages before they can analyze, or evaluate them. This research showed that students relied on content knowledge from the class in order to perform the data analysis tasks on the assessment instruments. In essence, this means that the assessment scenarios need to be paired to the content being covered in the class. This idea matches well with

Schafer's assertion (2011) that content needs to be aligned with assessment. An example, of what this could look like is shown in Appendix E.

After the timeline is revised and new assessments are created using the criteria above, the details of the assessments need to be improved. Cognitive interviews indicated there was a large amount of response error. Recall that response error occurs when subjects misunderstand the question. Willis (2005) recommends revising language in response to cognitive interviews. An example of how some of the assessments in this study might be modified is seen in Appendix F.

This research also had issues related to inter-rater reliability of scoring. This was caused by a poorly worded rubric that didn't have enough levels. Leighton (2011) recommends that rubrics have four or five levels whereas many questions in this research used only 3. Leighton also recommends that rubrics have clear language and expectations. The rubric used had some issues with clarity. Wiggins and McTighe (2005) recommend creating the rubric first or in conjunction with the assessments. Once the new rubric is created all existing assessments need to be revised to ensure students can adequately answer the questions given the scenario. An example of a potentially revised rubric is seen in Appendix G.

Once new assessments are created to match the criteria above, they must be checked for content validity. Much like the work done for this study each new assessment needs to be reviewed by an expert panel to ensure assessment reaches the threshold CVI value of 0.8 (Rubio, 2003). In this research, content validity was not enough. Messick (1995) recommends using three types of validity: content, construct and correlation.

Construct validity could be obtained using a series of interviews done before assessments were used. To obtain correlation validity, each assessment should be paired with a science inquiry experiment so that scores on the assessment could be validated versus scores on the science inquiry lab write up. See Appendix E for an example of how this might look. Additionally assessment instruments need to validated as being truly parallel so that gains between the pre-assessment and other assessments are related to student gains and not differences in assessment difficulty level. It should be noted that the assessments in Appendix F are not all the same difficulty level. If these assessments were to be used, they would need another revision to ensure they are parallel.

Pairing assessments with science inquiry addresses another issue that was found with this study. The initial design of this research involved giving students feedback on just the assessments. Unfortunately, fifty percent of the pre-assessments were blank. This study was modified and students were given evaluative feedback on a relevant assignment, the science inquiry lab. By aligning the assessments with science inquiry, students would get evaluative feedback on both. To facilitate evaluative feedback, a matrix of feedback could be created to give standardized individual feedback that students could understand. An example of this matrix is given in Appendix H. If there were issues with assessment participation, the science inquiry labs themselves could be used as the instruments for investigation. Once created, these assessments could be used as models to help teachers formatively assess skill development and content comprehension between

science inquiry labs or provide teachers with additional tools to improve data analysis skills.

It was not the intention of this study to examine the impact of literacy strategies on learning data analysis skills. Comments and actions of students during the cognitive interview indicate that literacy played a role. One student indicated that he did not know how to read science papers. Others struggled with vocabulary terms and tried to find meaning in context. Others commented that the more they interacted with material the more they understood it, which would imply that reciprocal teaching might also be useful. Because these assessments rely so heavily on text, it would be interesting to see the impact of different literacy strategies assessment scores. Fisher et al. (2002) indicate that use of literacy strategies aid student comprehension of difficult topics. If students don't understand the scenario in the assessment, they cannot engage in demonstrating their data analysis skills.

There were several questions that had high response error (Willis, 2005). Cognitive interviews provided insight on how question stems could be changed to improve understanding. It would be interesting to revise the assessments in light of these findings and see how students perform. In addition, the researcher could ensure assessments aligned with content so students could use their experiential knowledge from labs to support their claims.

Finally, motivation was not investigated as part of this study, but it many of the results seem to be tied to motivation. It would be good to have a measure of motivation as part of future research. This would help the researcher understand more fully how students approach tasks and why.

Implications to practice

The classroom needs to be established as a learning environment. This means creating an inviting space were teachers partner with students towards understanding. Learning goals should be created in advance and communicated to students. Formative assessment should be integrated into curriculum and clearly aligned with the learning goals. Feedback needs to be timely, goal oriented, and given before students get a "grade" on their work. Students need a chance to ask questions about this feedback and incorporate it into future work. Summative instruments need to be aligned with classwork and the goals to ensure students can demonstrate the knowledge they have learned.

Opportunities to provide students with autonomy, relatedness (purpose) and competence (mastery) should be sought out. One method to incorporate these ideas is through a series of science inquiry labs that begin as structured inquiry and slowly build towards open-ended inquiry. This approach to inquiry not only fosters autonomy, but also helps students develop the higher order cognitive skills required for college and career success. These experiences need the appropriate level of scaffolding to keep students in their ZPD without undermining higher order thinking.

Labs need to be tied to both interesting ideas and curriculum goals. This means creating fun experiences that tie back to the standards and having appropriate discussions that allow students to practice data analysis skills like scientific explanation and error analysis.

Although this research did not examine literacy strategies, literacy strategies seem to improve student understanding. Incorporating some of the seven defensible literacy strategies from Fisher et al.'s (2002) paper may help all students. Special attention needs to be paid to vocabulary terms. Students need time to construct definitions based on experiences or vocabulary terms may interfere with learning.

Assessment is not always straightforward. Adult content experts do not view questions in the same way as high school student. Students can often misinterpret questions and answer them in different ways. If there is a pattern of high numbers of students getting a certain question wrong on an assessment, it is worth taking time to understand what they think the question is asking. Learning why students are missing a problem can help guide the instruction. It may be that students do not understand the concepts, but it could also be that they don't get the question.

Citations

Anderson, L. W., Krathwohl, D. R., & Bloom, B. S. (2001). *A taxonomy for learning, teaching, and assessing.* New York, NY: Longman.

Baker, R. S., Corbett, A. T., & Koedinger, K. R., (2004) Learning to distinguish between representations of data: A cognitive tutor that uses contrasting cases. *Proceedings of the 6th international conference on Learning sciences.* Santa Monica, California. International Society of the Learning Sciences.

Bell, R. L., Smetana, L., & Binns, I. (2005). Simplifying inquiry instruction. *The Science Teacher*, *72*(7), 30-33.

Barr, S., Eslami, Z. R., & Joshi, R. M. (2012). Core strategies to support English language learners. *In The Educational Forum, 76*(1), 105-117.

Bennett, R. E., (2011) Formative assessment: a critical review. *Assessment in Education: Principles, Policy & Practice. 18*(1), 5-25.

Berg, C. A. R., Bergendahl, V. C. B., Lundberg, B., & Tibell, L. (2003): Benefiting from an open-ended experiment? A comparison of attitudes to, and outcomes of, an expository versus an open-inquiry version of the same experiment, *International Journal of Science Education*, *25*(3), 351-372

Berland, L. K., & Hammer, D. (2012). Framing for scientific argumentation. *Journal of Research in Science Teaching*, *49*(1), 68-94.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7-75.

Branan, D., & Morgan, M. (2009). Mini-lab activities: Inquiry-based lab activities for formative assessment. *Journal of Chemical Education*, *87*(1), 69-72.

Brophy, J. E. (2010). *Motivating students to learn*. New York, NY: Routledge.

Burns, M. K., Klingbeil, D., A., & Ysseldyke, J., (2010) The effects of technology-enhanced formative evaluation on student performance on state accountability math tests. *Psychology in the Schools. 47*(6). 582-91.

Carlson, M. O. B., Humphrey, G. E., & Reinhardt, K. S. (2003). *Weaving science inquiry and continuous assessment: Using formative assessment to improve learning*. Thousand Oaks, CA : Corwin Press.

Colburn, A. (2000). An inquiry primer. *Science Scope*, *23*(6), 42-44.

Corliss, S. B., Linn M. C., (2011) Assessment learning from inquiry science instruction. In G. Schraw and D. R. Robinson (Eds.) *Assessment of Higher Order Thinking Skills.* (pp 219-243). Charlotte, NC: IAP-Information Age Publishing, Inc.

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of educational research*, *58*(4), 438-481.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological bulletin*, *52*(4), 281-302.

Davila, K., & Talanquer, V. (2009). Classifying end-of-chapter questions and problems for selected general chemistry textbooks used in the United States. *Journal of Chemical Education*, *87*(1), 97-101.

Doige, C. A. (2012). E-Mail-based formative assessment: A chronicle of research-inspired practice. *Journal of College Science Teaching*, *41*(6), 32-39.

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

Dunbar, K. N., Fugelsang, J. A., & Stein, C. (2007). Do naïve theories ever go away? Using brain and behavior to understand changes in concepts. In *Thinking with data: 33rd carnegie symposium on cognition* (pp. 193-206).

Dunn, K. E., & Mulvenon, S. W., (2009). A critical review of research on formative assessment: The limited scientific evidence on the impact of formative assessment in education. *Practical Assessment, Research & Evaluation. 14*(7),

Fisher, D., Frey, N., & Williams, D. (2002). Seven literacy strategies that work. *Educational Leadership*, *60*(3).

Furtak, E. M., & Ruiz‐Primo, M. A. (2008). Making students' thinking explicit in writing and discussion: An analysis of formative assessment prompts. *Science Education*, *92*(5), 799-824.

Gray, K., Owens, K., Liang, X., & Steer, D. (2012). Assessing multimedia influences on student responses using a personal response system. *Journal of Science Education and Technology*, *21*(3), 392-402.

Herman, J., Linn, R., & Moss, F. (2013). On the road to assessing deeper learning: The status of smarter balanced and PARCC assessment consortia. CRESST Report 823. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Hofstein, A. (2004). The laboratory in chemistry education: Thirty years of experience with developments, implementation, and research. *Chem. Educ. Res. Pract.*, 5(3), 247-264.

Hofstein, A., & Lunetta, V. N. (1982). The role of the laboratory in science teaching: Neglected Aspects of Research. *Review of Educational Research*, *52*(2) 201-217.

Hofstein, A., Shore, R., & Kipnis, M. (2004). Providing high school chemistry students with opportunities to develop learning skills in an inquiry-type laboratory: A case study. *International Journal of Science Education*, *26*(1), 47-62.

Huddle, P. A., & White, M. D., (2000). Using a teaching model to correct known misconceptions in electrochemistry. *Journal of Chemical Education, 77*(1). 104-110

Hug, B., & McNeill, k. L., (2008). Use of first-hand and second-hand data in science: Does data type influence classroom conversations? *International Journal of Science Education, 30*(13). 1725-51.

Hume, A., Coll, R. K., (2009). Assessment of learning, for learning and as learning: New Zealand case studies. *Assessment in Education: Principles, Policy & Practice, 16*(3), 269-90.

Jang, H. (2008). Supporting students' motivation, engagement, and learning during an uninteresting activity. *Journal of Educational Psychology, 100*(4), 798.

Janzen, J. (2008). Teaching English language learners in the content areas. *Review of Educational Research, 78*(4), 1010-1038.

Karee, D. E., & Sean, M. W. (2009). A critical review of research on formative assessment: The limited scientific evidence of the impact of formative assessment in education. *Practical Assessment, Research & Evaluation, 14*(7), 1-11.

Kanari, Z. Millar, R., (2004). Reasoning from data: How students collect and interpret data in science investigations. *Journal of Research in Science Teaching, 41*(7), 748-69.

Keiler, L. S., (2007). Student's explanations of their data handling: Implications for transfer of learning. *International Journal of Science Education, 29*(2), 151-172.

Keys, C. W, Hand, B., Prain, V., & Collins, S. (1999). Using the science writing heuristic as a tool for learning from laboratory investigations in secondary science. *Journal of research in science teaching, 36*(10), 1065-84

Koedinger, K. R., McLaughlin, E. A., & Heffernan, N. T. (2010). A quasi-experimental evaluation of an on-line formative assessment and tutoring system. *Journal of Educational Computing Research*, *43*(4), 489-510.

Lee, S. E., Woods, K. J., & Tonissen, K. F., (2011) Writing activities embedded in bioscience laboratory courses to change students' attitudes and enhance their scientific writing. *Eurasia Journal of Mathematics, Science & Technology Education, 7*(3) 193-202.

Leighton, J. P., (2011) A cognitive model for the assessment of higher order thinking in students. In G. Schraw and D. R. Robinson (Eds.) *Assessment of Higher Order Thinking Skills.* (pp 151-181). Charlotte, NC: IAP-Information Age Publishing, Inc.

Lovett, M. C., & Chang N. M., (2007) Data-analysis skills: What and how are students learning? In M. Lovett and P. Shah (Eds.) *Thinking with Data.* (pp 293-318). Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

Marx, R. W., Blumenfeld, P. C., Krajcik, J. S., Fishman, B., Soloway, E., Geier,R., Tal. R. T. (2004). Inquiry-based science in the middle grades: Assessment of learning in urban systemic reform. *Journal of research in science teaching, 41*(10), 1063-80.

McIntosh, J., White, S., & Suter, R. (2009). Science sampler: Enhancing student understanding of physical and chemical changes. *Science Scope*,*33*(2), 54-58.

McNeill, K. L., & Krajcik, J., (2007) Middle school student's use of appropriate and inappropriate evidence in writing scientific explanations. In M. Lovett and P. Shah (Eds.) *Thinking with Data.* (pp 293-318). Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

McNeill, K. L., Lizotte, D. J., Krajcik, J., & Marx, R. W. (2006). Supporting students' construction of scientific explanations by fading scaffolds in instructional materials. *The Journal of the Learning Sciences*, *15*(2), 153-191.

Messick, S. (1990). *Validity of test interpretation and use*. Princeton, N.J. : Educational Testing Service

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741-749.

Miri, B., David, B. C., & Uri, Z. (2007) Purposely teaching for the promotion of higher-order thinking skills: A case of critical thinking. *Research in Science Education, 37*(4), 353-369.

Mullis, I. V., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., & Preuschoff, C. (2009). TIMSS 2011 assessment frameworks. Amsterdam, Netherlands: *International Association for the Evaluation of Educational Achievement.*

National Research Council (NRC). (1996). *3 science teaching standards. National science education standards*. Washington, DC: The National Academies Press.

National Research Council (NRC). (2000). *How people learn: Brain, mind, experience, and school: expanded edition*. Washington, DC: The National Academies Press.

National Research Council (NRC). (2011). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas.* Washington, DC: The National Academies Press.

Nyachwaya, J. M., Mohamed, A. R., Roehrig, G. H., Wood, N. B., Kern, A. L., & Schneider, J. L. (2011). The development of an open-ended drawing tool: An alternative diagnostic tool for assessing students' understanding of the particulate nature of matter. *Chem. Educ. Res. Pract.*, *12*(2), 121-132.

Overman, M., Vermunt, J. D., Meijer, P. C., Bulte, A. M., & Brekelmans, M. (2012). Textbook questions in context-based and traditional chemistry curricula analyzed

from a content perspective and a learning activities perspective. *International Journal of Science Education,* (ahead-of-print), 1-25.

Phelan, J. C., Choi, K., Niemi, D., Vendlinski, T. P., Baker, E. L., & Herman, J. (2012). The effects of POWERSOURCE© assessments on middle-school students' math performance. *Assessment in Education: Principles, Policy & Practice*, *19*(2), 211-230.

Pink, D. (2009). *Drive: The surprising truth about what motivates us.* New York, NY: Penguin.

Rubio, D. M., Berg-Wegner, M., Tebb, S. S, Lee, E. S., & Rauch, S. (2003) Objectifying content validity: Conducting a content validity study in social work research. *Social Work Research, 27*(2), 94-104.

Ruiz-Primo, M. A., & Furtak, E. M. (2007). Exploring teachers' informal formative assessment practices and students' understanding in the context of scientific inquiry. *Journal of Research in Science Teaching, 44*(1) 57-84.

Ruiz-Primo, M. A., Li, M., Tsai, S. P., & Schneider, J. (2010). Testing one premise of scientific inquiry in science classrooms: Examining students' scientific explanations and student learning. *Journal of Research in Science Teaching*, *47*(5), 583-608.

Ruiz-Primo, M. A., & Shavelson, R. J. (1996). Rhetoric and reality in science performance assessments: An update. *Journal of Research in Science Teaching*, *33*(10), 1045-1063.

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American psychologist*, *55*(1), 68.

Ryan, R. M., & Niemiec, C. P (2009). Self-determination theory in schools of education: Can an empirically supported framework also be critical and liberating*? Theory and Research in Education. 7*(2), 263-72.

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

Schafer, W. D., (2011) Aligned by design: A process for systematic alignment of assessments to educational domains in G. Schraw and D. R. Robinson (Eds.) *Assessment of Higher Order Thinking Skills.* (pp 395-418). Charlotte, NC: IAP-Information Age Publishing, Inc.

Schauble, L., Glaser, R., Duschl, R.A., Schulz, S., & John, J. (1995). Students' understanding of objectives and procedures of experimentation in the science classroom. *The Journal of the Learning Science*, *4*(2), 131–166.

Shah, P., & Hoeffner, J. (2002). Review of graph comprehension research: Implications for instruction. *Educational Psychology Review, 14*(1), 47–69.

Shavelson, R. J., Young, D. B., Ayala, C. C., Brandon, P. R. Furtak, E. M, Ruiz-Primo, M. A, Tomita, M. K., & Tin, Y. (2008) On the impact of curriculum-embedded formative assessment on learning: A collaboration between curriculum and assessment developers, *Applied Measurement in Education*, *21*(4) 295-314.

Shute, V. J. (2008). Focus on formative feedback. *Review of educational research*, *78*(1), 153-189.

Van Der Stuyf, R. R. (2002). Scaffolding as a teaching strategy. *Adolescent learning and development*, 2-13.

Tama, M. C., & Haley, A. M. (2007) *Reading and writing in the content areas: Practical strategies.* Dubuque, IA: Kendall Hunt.

Torrance, H., (2007) Assessment *as* learning? How the use of explicit learning objectives, assessment criteria and feedback in post-secondary education and training can come to dominate learning. *Assessment in education,* 14(3) 281-294.

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

Vital, F. (2011). Creating a positive learning environment with the use of clickers in a high school chemistry classroom. *Journal of Chemical Education*, *89*(4), 470-473.

Vygotsky, L. S., (2011) The dynamics of the schoolchild's mental development in relation to teaching and learning. (A. Kozulin, Trans.) *Journal of Cognitive Education and Psychology, 10*(2) 198-210. (Original work published 1935).

Walpuski, M., Ropohl, M., & Sumfleth, E. (2011). Students' knowledge about chemical reactions–development and analysis of standard-based test items. *Chem. Educ. Res. Pract., 12*(2), 174-183.

Wheeler, L., & Bell, R. (2012). Open-ended inquiry. *Science Teacher*, *79*(6), 32-39.

Wiggins, G. P., & McTighe, J. A. (2005). *Understanding by design*. Alexandria, Va.: Association for Supervision and Curriculum Development.

Willis, G. B. (2005). *Cognitive interviewing: A tool for improving questionnaire design.* Thousand Oaks, CA: SAGE Publications, Incorporated.

Yeh, S. S. (2001). Tests worth teaching to: Constructing state-mandated tests that emphasize critical thinking. *Educational Researcher, 30*(9), 12-17.

Appendix A: Instructional support materials

Water Fluoridation

Brief summary from the internet

- Adding Fluoride to the city's water is a good idea. Fluoridation reduces tooth decay by 20-40% (ADA). For every dollar spent on fluoridation, $38 dollars are saved in dental care. Low levels of fluoridation are safe. If you compare states that fluoridate their water versus those that do not, there is no evidence of increased cancer or other diseases in states that fluoridate water. A 2008 study on Fluoride's effect on osteoporosis found that daily doses of up to 20 mg fluoride significantly increased bone mineral density and reduced fracture risk. Fluoridation decreases tooth decay with no negative health issues and tooth decay is linked to other diseases. We should add fluoride to our water.
- Adding fluoride to our city's water is a bad idea. It does not help prevent tooth decay. In New Zealand areas with fluorinated drinking water and those without have shown the same reduction in tooth decay. High levels of fluoride consumption leads to dental fluorosis (dark streaks, brown spots and pitting of the enamel). Fluoride levels of 0.2mg/L or more have been linked to a 55% reduction in salmon migration in the pacific northwest. One study on rats and mice indicated higher levels of benign bone tumors in animals with a high dose of fluoride. Water fluoridation will cost Oregon tax payers $5 million dollars. Portland water is really good and fluoridation could affect the taste. The benefits of fluoridation may not be real. Over fluoridation can cause real health problems and hurt the environment. We should not add Fluoride to our water

Appendix B: Science Inquiry lab packet

Science Inquiry Lab - This lab is worth 5% of your grade and must be turned in.

Background and Hypothesis

What reaction are you doing? What factor are you testing? Why are you testing this? What do you think will happen? What is your hypothesis? What do you think will happen and why?

Procedures

What did you do? Describe what happened.

Results

What are your results? What did you see? What did you measure?

Analysis

What do your results mean? (don't forget to include your claim evidence and reasoning)If you could do another test, what would you test and how would you do it? What do you think could have affected your lab?

Example one

Background and Hypothesis

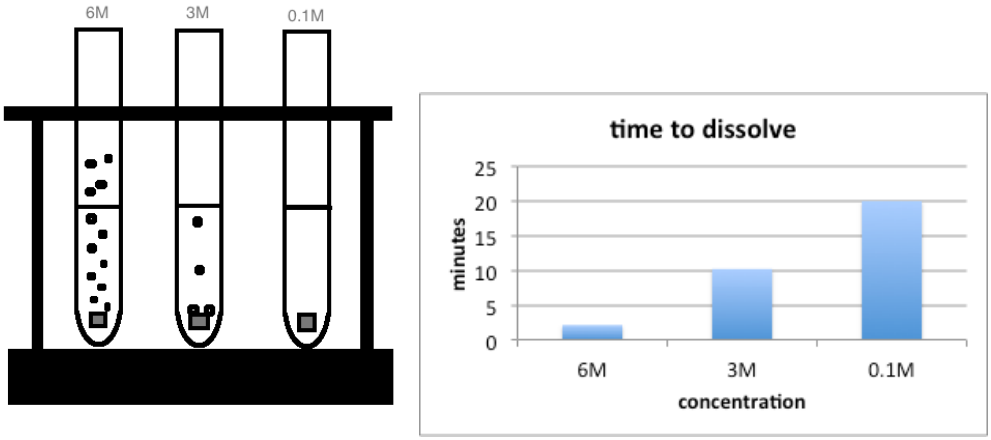
Hydrochloric acid reacts with zinc to produce hydrogen bubbles. I want to slow this reaction down. I believe that HCl will react slower with Zn as we increase concentration because the solution will get too crowded.

Procedures

We got 3 clean dry test tubes and put them in a test tube rack. We added 5 mL of 6M HCl (high concentration) to one test tube, 5mL of 3M HCl (medium concentration) to a second test tube, and 5mL of 0.1M HCl (low concentration) to a third test tube. We found three pieces of Zn that appeared to be approximately the same size. I added one piece of zinc to each of the three test tubes.

Results

Below is a diagram of what we saw



The 6M HCl bubbled quickly and the Zn seemed to dissolve within a few minutes. The 3M HCl bubbled less and it took longer for the Zn to dissolve. We got a little bored. We got tired of waiting for the 0.1M HCl to do anything so we just said it took 20 minutes.

Analysis

For example, we thought that a higher concentration of HCl would react more slowly than a low concentration of HCl. Instead the higher the concentration of HCl, the faster the reaction happened, which contradicted our hypothesis. The highest concentration, 6M, reacted most rapidly releasing a steady stream of bubbles. The Zn dissolved quickly indicating that the reaction was happening. The intermediate concentration, 3M, had bubbles appear on the surface, but only released a few. It took a lot longer for all the Zn to dissolve. The low concentration didn't appear to react at all.

The concentration affected the reaction rate because there were more molecules of HCl available to react with the Zn. These molecules needed to encounter the surface of the Zn in order for the single replacement reaction shown below to happen

$$2HCl + Zn \rightarrow ZnCl_2 + H_2$$

The next experiment I would like to try is to see what effect temperature would have on the reaction. I would like to use 3M HCl and test it at a high temperature (in a boiling water bath), room temperature, and cold temperature (ice water bath)

I could have measured the mass of everything in the reaction so I could determine exactly how much hydrogen was created. (design/measurement). I could have had different levels of acid concentrations to graph this reaction better (design). I could have drawn a line through my data points and used that to predict reaction rates at other concentrations. I might have measured the HCl incorrectly or spilled some on the table and this would cause my reaction to be off. (execution)

Example two

Background and Hypothesis

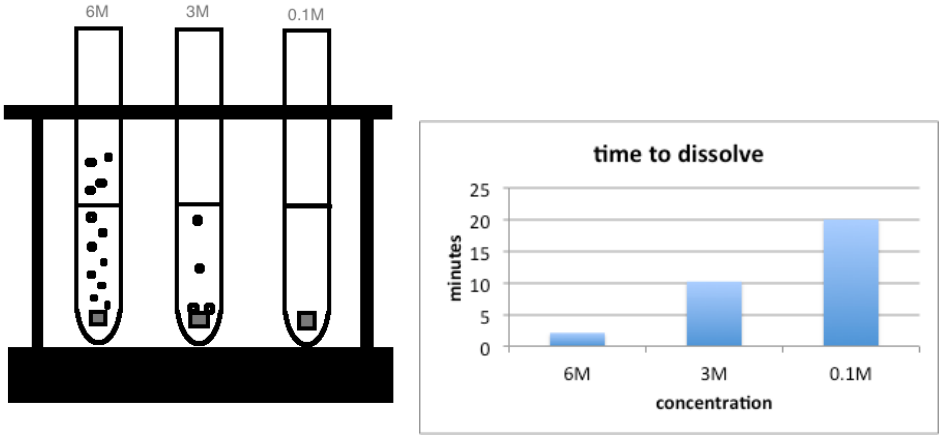
I believe that HCl will react slower with Zn as we increase concentration.

Procedures

We added 5 mL of 6M HCl, 5mL of 3M HCl, and 5mL of 0.1M HCl to three different test tubes.. I added one piece of zinc to each of the three test tubes.

Results

Below is a diagram of what we saw



The 6M HCl bubbled quickly and the Zn seemed to dissolve within a few minutes. The 3M HCl bubbled less and it took longer for the Zn to dissolve. We got a little bored. We got tired of waiting for the 0.1M HCl to do anything so we just said it took 20 minutes.

Analysis

The 6M HCl reacted faster than anything else. It might have done this because we messed up. My partner was joking around with me and told me that the test tubes were in a different order, but I didn't really know. The teacher came over and told us that it looked like one of them was the 6M and so I believed her. The bubbles were cool I guess, but I liked the reactions that changed color better. We've done too many reactions with baking soda and vinegar. Maybe we could blow something up or burn something.

The zinc disappeared while I was in the bathroom and my partner didn't notice so I made a guess about the times. I think it happened this way because molecules move faster when they are hot and they need enough energy to react. Below is the reaction we had in class.

$$HCl + Zn \rightarrow ZnCl_2 + H_2$$

The next experiment I would like to burn something. I think maybe we could see how fast different paper airplanes burned. Maybe we could fold them different ways or use different colors.

I think we could have paid attention better to measuring and what was happening in the lab. Maybe if my partner had taken some notes it would have helped I guess.

Scoring rubric for lab

Scientific Inquiry Scoring Guide:

section	Not proficient	proficient
Hypothesis and background	Hypothesis cannot be tested Background is unclear or does not support hypothesis	Hypothesis is testable Background establishes reason for experiment
Experimental design	Procedure has some errors or had to follow. Experiment will not provide enough data to test hypothesis	Logical procedure easily followed Experiment will give enough data to test hypothesis
Collecting and Presenting data	Data collected inconsistent with plan, missing units and poorly organized	Data collected consistent with plan, contains appropriate units and organized in table or graph
Analyzing and interpreting results	Conclusion not supported by evidence, no real reasoning and communicates using unscientific terms	Valid conclusion based on data. Addresses hypothesis. Evidence provided and connected to claim with reasoning. Uses appropriate scientific terms

Appendix C: Grading Rubric for Student Assessments

Question 1 – graph interpretation

0 – blank or unrelated answer

1 – Student inaccurately references graph. There may be a small misreading of the graph

2 – Student accurately read value from the graph

Question 2 drawing conclusions (sum of the following three scores)

Claim

0 – no claim or claim is unrelated to scenario

1 – claim inaccurately or vaguely references scenario

2 – Claim is explicitly stated, related to the scenario and supported by the data

Evidence

0 – no data cited or data is unrelated to scenario

1 – Student inaccurately cites scenario or vaguely references scenario

2 – Student accurately uses at least two pieces of evidence from the scenario to support claim

Reasoning

0 – no reasoning given, reasoning does not relate to scenario or reasoning is illogical

1 – Student reason contains some flaws in logic or inaccurately represents science

2 – reasoning supporting claim makes sense and demonstrates understanding

Question 3 – Drawing inferences and recognizing patterns (sum of the following 2 scores)

Making predictions

0 – no prediction made or prediction seems like wild guess

1 – Student's prediction is inaccurate or only vaguely supported by the data

2 – Student's prediction accurately represents the data in the scenario

Explaining prediction

0 – no explanation or explanation does not make sense

1 – explanation based on scenario, but is inaccurate or only vaguely related to the scenario

2 – Student's explanation accurately represents science

Question 4 – Hypothesis testing

0 – question blank or data required seems unrelated to scenario

1 – data required would not answer key question, but indicates student understands the idea of hypothesis testing

2 – data required would verify claim

Question 5 - error analysis

0 - blank or answer does not make sense

1 - identifies one limitation for data in depth or two superficially (design, representation, execution or measurement)
2 - identified three of the limitations of data with at least one in depth (design, representation, execution or measurement)
3 - identifies all four limitations of data with at least one in depth or identifies 3 limitations of data with at least 2 in depth.
4 - identifies all four limitations of data with at least 3 described in depth

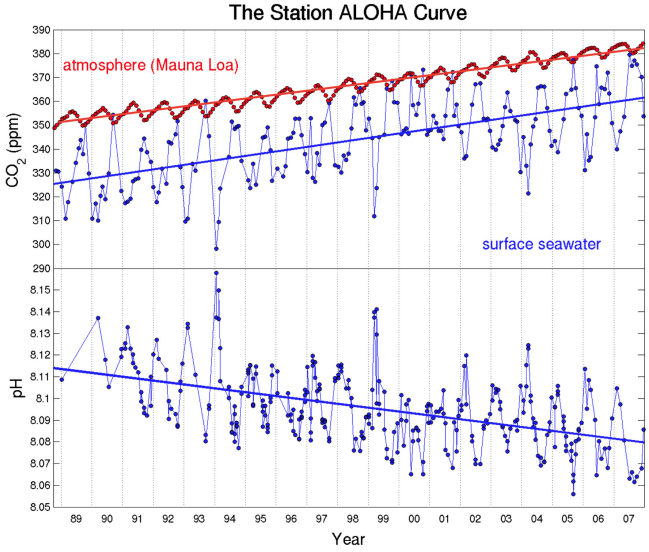
Appendix D: student assessments
Pre-assessment

pH is a measure of the acidity of a solution. There are multiple ways to determine pH. A pH meter will give you a number. An indicator paper or indicator solutions will change color at a certain pH. On the pH scale, the lower the number, the more acidic the solution.

Different reactions change the pH of a solution. When carbon dioxide (CO_2) dissolves in water, a small amount of it reacts with the water to make carbonic acid. This can make the solution more acidic.

Here is the reaction that makes carbonic acid: $CO_2 + H_2O \rightarrow H_2CO_3$

Below is a measure of acidity versus time at a specific place in the Pacific Ocean (bottom graph). There is also a plot of the dissolved CO_2 in the surface water (top graph). The solid lines are trend lines to help you see where the data is going.



http://www.sciencebuzz.org/buzz-tags/ocean-acidification

Vocabulary
Carbonic acid – an acid made when carbon dioxide dissolves in water.
pH – a measure of the acidity of solution
pH meter – device to measure pH
indicator paper – paper that changes color to indicate the pH of a solution
indicator solution – liquid you add to a solution that changes color at specific pH
surface seawater – water close to the ocean surface

Graph comprehension
In what year did the pH first drop below 8.07 at this location?

Graph interpretation
In your own words explain what is happening in this graph.

Claim –

Evidence –

Reasoning –

Drawing conclusions from data
If this trend continues, predict whether you think the pH will ever get below 8.05. How long do you think that would take?

Explain why you think the pH will or will not get to 8.05.

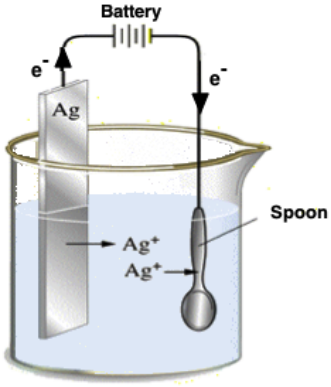
Your partner says by August this year the pH will be below 7.5. What data would you need to prove your partner wrong?

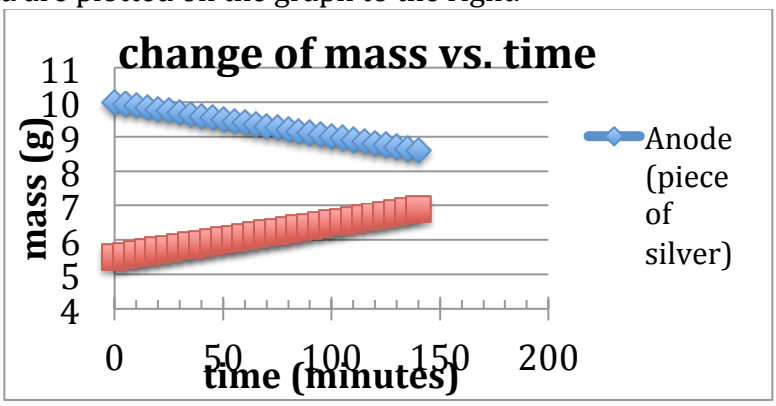
What data would you need to determine if these measurements were accurate?

Intermediate assessment

Electroplating is a process where you use electricity to deposit metal on a conductive surface. A battery is hooked up to a metal and takes electrons from one piece of metal and moves them to a second piece of metal. As these electrons are removed from the first piece of metal, the atoms on the surface are oxidized (lose electrons). This causes them to become ions that dissolve in the liquid. As electrons build up in the other piece of metal, ions in solution can be reduced (gain electrons) and deposit as solid metal on the surface. The rate at which metal is deposited is determined by how much electricity is supplied and how many contaminants are in the solution.

Say that you found some old spoons that were once silver plated. You found out that the spoons are worth more money if you can re-plate them with silver. As an experiment, you try the set-up on the left. To monitor your progress, you determine the mass of the spoon and the mass of the silver bar every minute during the silver-plating process. These data are plotted on the graph to the right.





Terms:

Anode – electrode where oxidation occurs. In this case the silver is the anode.
Cathode – electrode where the reduction occurs. In this case the spoon is the cathode
Electrochemical process – a reaction where electrons are transferred from one atom, molecule, or ion to another atom, molecule, or ion.
Silver plating – an electrochemical process where a thin layer of silver metal is deposited on the surface of another metal
Oxidation – when an atom, molecule or ion loses an electron

Graph comprehension **name:**

How long does it take for the spoon to weigh 6 grams?

Graph interpretation

In your own words explain what is happening in this graph.

Claim -

Evidence -

Reasoning -

Drawing conclusions from data

Predict whether the spoon will ever get to the point where it weighs 7.5 grams? If you think it will, how long will it take?

Explain why you think spoon will or will not get to 7.5 grams.

You set this up at 10am and your partner says the silver bar will be gone before tomorrow. What data would you need to prove your partner wrong?

What data would you need to determine if these measurements were accurate?

Post assessment

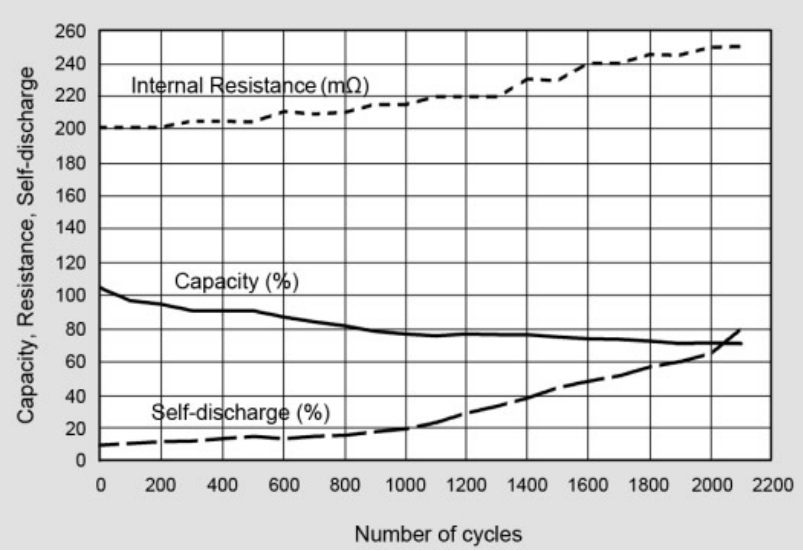
A battery is a type of electrochemical cell. The anode and cathode are labeled as the negative and positive parts of the battery. When a battery is put into a device (like a phone), electrons travel through the device (your phone) from the anode to the cathode. This powers up the device allowing it to run.

Below is a graph showing how charge capacity, internal resistance and self-discharge all vary with the number of times the battery has been recharged. Capacity is the amount of charge that the battery can hold. It is measured versus the capacity that is claimed for the battery. This is because batteries quickly lose capacity at first. As a result, the capacity of a new battery is slightly above the capacity that is claimed.

Internal resistance indicates how much the battery will heat up when it is being used or charged.

Self-discharge indicates how much charge the battery will lose when not in use. When your phone is new, you usually don't see a difference. If you have an old phone, you might notice a giant drop in charge if it is not plugged in overnight.

Below is a graph of the charge capacity vs. the number of times the battery has been recharged for an ultra-high-capacity nickel-cadmium battery. In the battery industry the number of times the battery has been charged is called the number of cycles. To measure these numbers, batteries are charged and then tested with a multimeter to determine the internal resistance, and the charge. They are left overnight and tested again to determine the self-discharge level.



http://batteryuniversity.com/learn/article/battery_performance_as_a_function_of_cycling

Terms:

Capacity – energy stored in a battery. Usually defined as the number of hours that a battery can provide a specified voltage – think of it as how long can the battery run your phone

Self discharge – the percentage of capacity that is lost when the battery sits on a shelf without being plugged in. For example, if you leave your phone turned off and then turn it on later does it still have the same level of charge as when you turned it off?

Internal resistance – This is a measure of battery life. It indicates how much the battery will heat up when it is being charged or being used.

Note: capacity and self discharge are not dependent on each other. They are both affected by the number of cycles

Graph interpretation
If you needed a battery to have a capacity of at least 80%, how many cycles could you go through before you had to replace the battery?

Drawing conclusions and scientific claims
In your own words explain what is happening in this graph.

Claim –

Evidence –

Reasoning -

Drawing inferences and recognizing patterns
Predict whether the battery will ever get to the point where the capacity is less than 50%. If you think that it will get there, how long will it take?

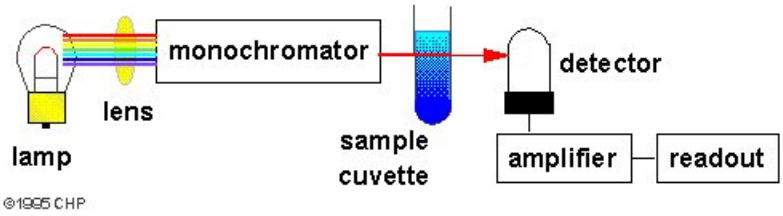
Explain why you think the battery will or will not get to 50%

Your partner says the self discharge rate will hit 100% before 3000 cycles. What data would you need to prove your partner wrong?

5. Identifying sources of error
What data would you need to determine if these measurements were accurate?

Think-aloud assessment

A UV-VIS absorption spectrometer uses light to detect small changes in color within a solution. Below is the simplified setup. Light goes through a sample and the instrument detects what colors are absorbed. One way scientists use this is to see how fast a solution is changing color.



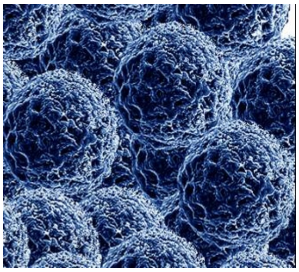
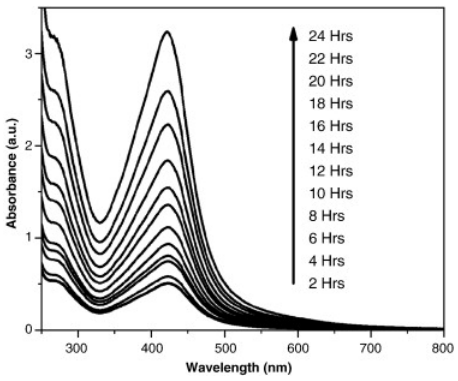
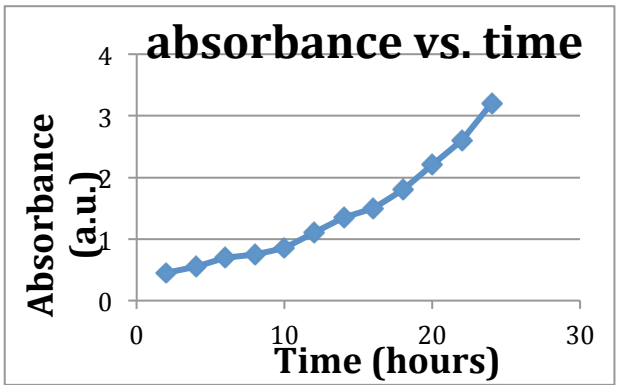


Image from http://www.chemistry.adelaide.edu.au/external/soc-rel/content/uv-vis.htm image of silver nanoparticles

Silver nanoparticles are tiny bits of silver that are smaller than the width of a human hair. (a magnified image of them is shown above on the right) One group of scientists recently used a UV-VIS absorption spectrometer to see how fast they could grow silver nanoparticles with a fungus. Silver nanoparticles absorb a specific color (referred to as a specific wavelength of light). The graph below on the left indicates that this wavelength of light is 420nm, which is purple/violet. Scientists measure how much 420nm light is absorbed to determine how many silver nanoparticles have formed. A higher absorbance = more nanoparticles. To make it easier to read, we have added the graph on the right, which just plots the absorbance of 420nm with time.





http://www.sciencedirect.com/science/article/pii/S1549963409000914

Terms:

UV-Vis – the spectrum of light that includes ultraviolet all the way through visible light

Spectrometer – a device for measuring wavelengths of light

Monochrometer – a special filter used to control the wavelength of light

Sample cuvette – glass test tube with certain dimensions that contains your liquid sample

Detector – device to measure light intensity
Amplifier – helps the detector to measure light intensity
Light transmission – how much light passes through something. For example windows covered in dirt do not transmit as much light as clean windows.
Nanoparticle – a particle that is smaller than the width of a human hair (measured in nanometers)
Absorbance – how much light is absorbed by a solution

Graph comprehension
When does the peak at 420nm reach an intensity of 2?

Graph interpretation
In your own words explain what is happening in this graph.

Claim

Evidence

Reasoning

Drawing conclusions from data
Assume that the creation of silver nanoparticles continues at the same rate. Predict whether the absorbance will ever reach 4. If you think it will, how long will it take?

Explain why you think the absorbance will or will not ever reach 4

Your partner says he has some fungus in his back yard that could make silver nanoparticles. What data would you need to prove your partner wrong?

What data would you need to determine if these measurements were accurate?

Appendix E: Revised Assessment schedule

Assessment	Current subject	New subject	Paired inquiry	Type inquiry
Pre-assessment	Ocean pH	Same	pH	Structured
Intermediate	Electroplating	Solubility	Solubility	Guided
Final	Battery charging	Reaction rates	Reaction Rates	Open-ended
Think-aloud	Silver nanoparticles	Battery charging	Electrochemical cell	Open-ended

Brief summary of inquiry

pH inquiry

Students are asked to explore the relationship between concentration and pH. They are given two methods to measure pH: universal indicator strips and cabbage juice. Each student group is given a different set of 3 liquids. Each student group tests the pH of their solutions, dilutes the solutions and test again. The groups must dilute their solution again and test a third time. They plot the pH versus amount of water added. As a class the students discuss what they think is happening when water is added and how this affects pH. What is the relationship between pH and concentration based on all the examples in class?

Solubility

The recipe to make rock candy involves dissolving 4 cups of sugar in 2 cups of water. Students are asked to determine the optimal temperature for dissolving 2 cups of granulated sugar in 1 cup of water. They must make the water dissolve that amount within 5% error, even when excess sugar is added. Each group is given different amounts of sugar. No students are given the right proportions to test 2 cups of sugar in 1 cups of water. They must come up with a procedure to provide them enough data to make an educated guess. Each group's prediction will be tested by putting 3 cups of sugar in 1 cup of water at that temperature. The sugar water will be decanted off and students will receive a grade based on the weight of the left over sugar.

Reaction Rates

Students are given a choice of reactions (Diet coke and Mentos or starch and iodine clock reaction) Students are asked to pick one factor that can affect reaction rate. (temperature, inhibitor/catalyst, concentration). Students must create a set of data that will allow the to predict what will happen when the teacher reproduces the experiment with slightly different conditions. For example, The student is looking at how different numbers of mentos affect reaction rate. The student tests 1, 3 and 5 mentos will react. They must pick a number of mentos that they didn't use. For example, the student may pick 4 mentos. The student must predict what will happen. The student will be

graded based on how close the student prediction matches the teacher's results, when the teacher puts 4 mentos into a diet coke.

Electrochemical cell

Students are asked who can create the best battery. They are given different sizes, types and shapes of metals with different electrochemical potentials. They are given different electrolytes and different concentrations of electrolytes. Batteries will be judged based on power output, how long they generate power and cost. They must predict what will happen if one of the items is replaced. For example the students must predict what will happen if the electrolyte is replaced with another. They must generate enough data to make this prediction without running the experiment. They are graded based on how well their prediction matches reality.

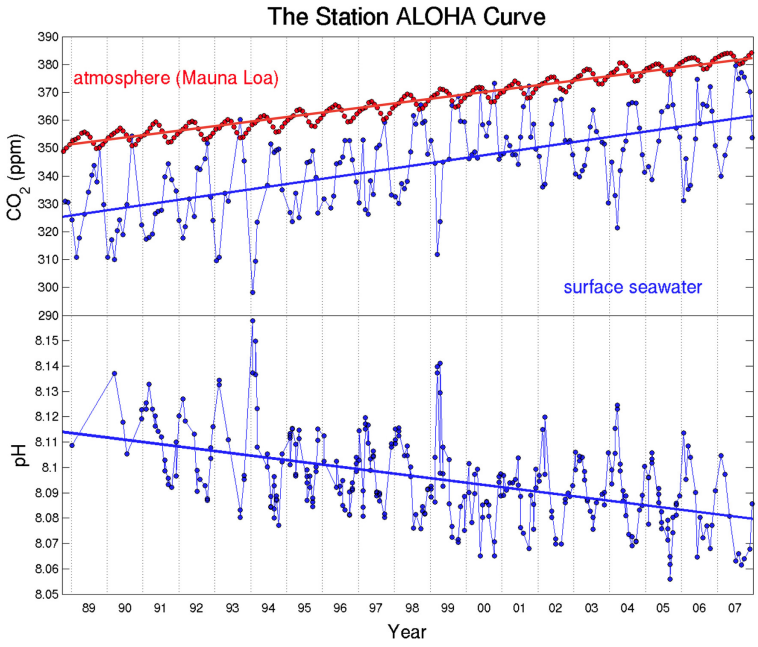
Appendix F: Revised student assessments
Pre-assessment

pH is a measure of the acidity of a solution. There are multiple ways to determine pH. A pH meter will give you a number. An indicator paper or indicator solutions will change color at a certain pH. On the pH scale, the lower the number, the more acidic the solution.

Different reactions change the pH of a solution. When carbon dioxide (CO_2) dissolves in water, a small amount of it reacts with the water to make carbonic acid. This can make the solution more acidic.

Here is the reaction that makes carbonic acid: $CO_2 + H_2O \rightarrow H_2CO_3$

Below is a measure of acidity versus time at a specific place in the Pacific Ocean (bottom graph). There is also a plot of the dissolved CO_2 in the surface water (top graph). The solid lines are trend lines to help you see where the data is going.



http://www.sciencebuzz.org/buzz-tags/ocean-acidification

Vocabulary

Carbonic acid – an acid made when carbon dioxide dissolves in water.
pH – a measure of the acidity of solution
pH meter – device to measure pH
indicator paper – paper that changes color to indicate the pH of a solution
indicator solution – liquid you add to a solution that changes color at specific pH
surface seawater – water close to the ocean surface

Reading a graph
In what year did the pH first drop below 8.07 at this location?

Scientific explanation
Read the scenario and look at the graph. Using your own words, explain what this research means. Including the following:

Claim (What do you think this research is saying?)–
Evidence (What evidence can you cite from the scenario or graph to support your claim?)–
Reasoning (Using what you know about pH and what is written in this scenario, explain how this evidence supports your claim and what is happening chemically)–

Recognizing patterns and making inferences
If you went to Aloha station and measured a value of 8.06 today, do you think this would indicate the trend is continuing or that the trend has changed? Using what you can infer from the pattern of data shown, explain why your measurement shows the trend is continuing or changing

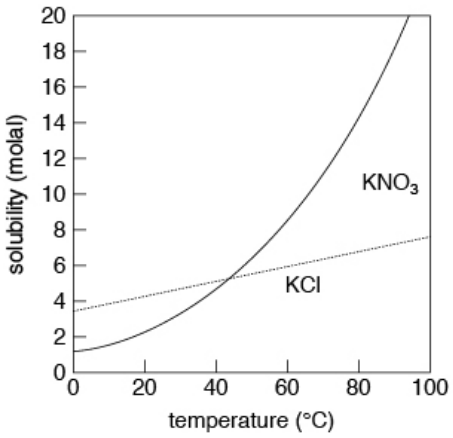
Hypothesis testing
Your partner says by August this year the pH will be below 7.9. Plan an experiment to test this hypothesis. What data would you need to prove your partner right or wrong?

Error analysis
What are the potential sources of error? What additional information would you need to determine if the initial researcher's data was correct?

Intermediate assessment

Many sports drinks like Gatorade contain a variety of salts dissolved in sugar water. Each of these salts can have a different solubility at a different temperature. Solubility is defined as the amount of solute that will dissolve in a solute and is determined by adding solute to a solvent until no more will dissolved. This was done in the graph below. For example, in the graph below the dotted line indicates how many moles of KCl can be dissolved in 1,000g of water at a specific temperature

Different ions change the solubility of compound. In the graph below potassium nitrate (KNO_3) has a higher solubility at 80°C than potassium chloride (KCl). Since both salts contain potassium, the difference in solubility is based on how well nitrate dissolves vs. chloride at this temperature.



Vocabulary
Solubility
Solute
Solvent -
Molal – Moles of solute dissolved in 1,000 grams of solution

Reading a graph
According to the graph, what is the solubility of KCl at 25°C?

Scientific explanation
Read the scenario and look at the graph. Using your own words, explain what this research means. Including the following:

Claim (What do you think this research is saying?)–
Evidence (What evidence can you cite from the scenario or graph to support your claim?)–
Reasoning (Using what you know about pH and what is written in this scenario, explain how this evidence supports your claim and what is happening chemically)–

Recognizing patterns and making inferences

do you think this would indicate the trend is continuing or that the trend has changed? Using what you can infer from the pattern of data shown, explain why your measurement shows the trend is continuing or changing

Hypothesis testing

Plan an experiment to test this hypothesis. What data would you need to prove your partner right or wrong?

Error analysis
What are the potential sources of error? What additional information would you need to determine if the initial researcher's data was correct?

Think-aloud

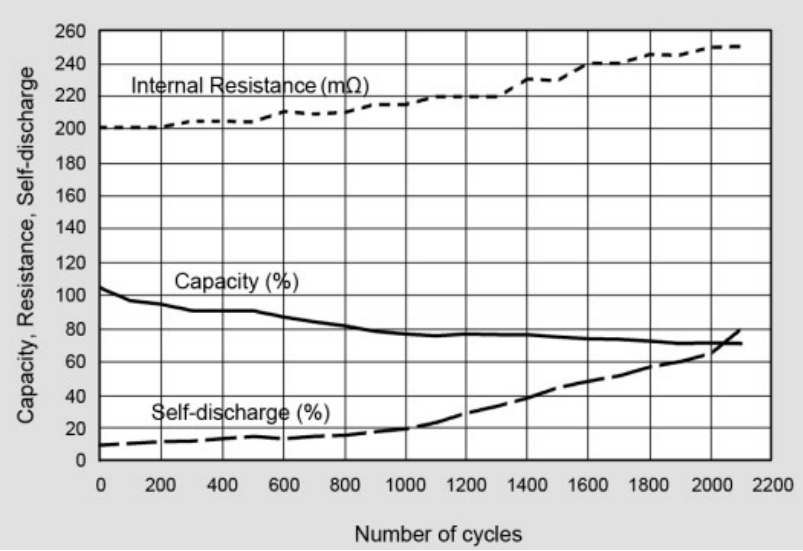
A battery is a type of electrochemical cell. The anode and cathode are labeled as the negative and positive parts of the battery. When a battery is put into a device (like a phone), electrons travel through the device (your phone) from the anode to the cathode. This powers up the device allowing it to run.

Below is a graph showing how charge capacity, internal resistance and self-discharge all vary with the number of times the battery has been recharged. Capacity is the amount of charge that the battry can hold. It is measured versus the capacity that is claimed for the battery. This is because batteries quickly lose capacity at first. As a result, the capacity of a new battery is slightly above the capacity that is claimed.

Internal resistance indicates how much the battery will heat up when it is being used or charged.

Self-discharge indicates how much charge the battery will lose when not in use. When your phone is new, you usually don't see a difference. If you have an old phone, you might notice a giant drop in charge if it is not plugged in overnight.

Below is a graph of the charge capacity vs. the number of times the battery has been recharged for an ultra-high-capacity nickel-cadmium battery. In the battery industry the number of times the battery has been charged is called the number of cycles. To measure these numbers, batteries are charged and then tested with a multimeter to determine the internal resistance, and the charge. They are left overnight and tested again to determine the self-discharge level.



http://batteryuniversity.com/learn/article/battery_performance_as_a_function_of_cycling

Terms:

Capacity – energy stored in a battery. Usually defined as the number of hours that a battery can provide a specified voltage – think of it as how long can the battery run your phone

Self discharge – the percentage of capacity that is lost when the battery sits on a shelf without being plugged in. For example, if you leave your phone turned off and then turn it on later does it still have the same level of charge as when you turned it off?

Internal resistance – This is a measure of battery life. It indicates how much the battery will heat up when it is being charged or being used.

Note: capacity and self discharge are not dependent on each other. They are both affected by the number of cycles

Reading a graph
If you needed a battery to have a capacity of at least 80%, how many cycles could you go through before you had to replace the battery?

Scientific explanation
Read the scenario and look at the graph. Using your own words, explain what this research means. Including the following:

Claim (What do you think this research is saying?)–
Evidence (What evidence can you cite from the scenario or graph to support your claim?)–
Reasoning (Using what you know about pH and what is written in this scenario, explain how this evidence supports your claim and what is happening chemically)–

Recognizing patterns and making inferences
If you measured the capacity of a battery after 1000 cycles and got a value that was 82%, do you think this would indicate the battery was similar to the one representedby the graph above or different? Using what you can infer from the pattern of data shown, explain why your measurement shows the trend is continuing or changing

Hypothesis testing
Your partner has a hypothesis that a battery will no longer function when the self-discharge % is the same as the capacity %. Plan an experiment to test this hypothesis. What data would you need to prove your partner right or wrong?

Error analysis
What are the potential sources of error? What additional information would you need to determine if the initial researcher's data was correct?

Appendix G: Revised Grading Rubric

Question 1 – graph interpretation
0 – blank or unrelated answer
1 – Student inaccurately references graph. There may be a small misreading of the graph
2 – Student accurately read value from the graph

Question 2 scientific explanation (sum of the following three scores)
Claim
0 – no claim or claim is unrelated to scenario
1 – Claim vaguely related to scenario, unclear and contains inaccuracies
2 – Claim is related to the scenario, but vague and contains some inaccuracies
3 – Claim is explicitly stated, related to the scenario, but contains some inaccuracies
4 – Claim is explicitly stated, related to the scenario and supported by the data

Evidence
0 – no data cited or data is unrelated to scenario
1 – Student inaccurately cites scenario or vaguely references scenario
2 – Student uses one or two pieces of data but does not seem to support claim
3 – Student uses one relevant piece of data to support claim and hints to comparison (
4 – Student accurately uses at least two pieces of evidence from the scenario to support claim or and talks about relationship between data

Reasoning
0 – no reasoning given, reasoning does not relate to scenario or reasoning is illogical
1 – Student reason contains major flaws in logic and does not support claim
2 – Student reasoning contains flaws in logic, vaguely supports claim or uses antidotal evidence to support claim.
3 – Student reasoning contains minor flaws in logic, demonstrates some understanding of the science but supports claim
4 – Student reasoning supports claim, makes sense and demonstrates understanding of the science

Question 3 – Drawing inferences and recognizing patterns (sum of the following 2 scores)
Making predictions
0 – no prediction made or prediction seems like wild guess
1 – Student's prediction is inaccurate or only vaguely supported by the data
2 – Student's prediction accurately represents the data in the scenario

(Continued on next page)

Explaining prediction
0 – no explanation or explanation does not make sense
1 – explanation based on scenario, but is inaccurate or only vaguely related to the scenario
2 – Student's explanation demonstrates some understanding of science but contains major flaws
3 – Student's explanation demonstrates some understanding of science but contains minor flaws
4 – Student's explanation accurately represents science

Question 4 – Hypothesis testing
0 – question blank or data required seems unrelated to scenario
1 – data required would not answer key question
2 – Data indicated could answer key question, but it is unclear how it will be used
3 – Student identifies data needed to verify claim, but testing plan has flaws
4 – Planned experiment would verify claim

Question 5 - error analysis
0 - blank or answer does not make sense
1 - identifies one limitation for data in depth or two superficially (design, representation, execution or measurement)
2 - identified three of the limitations of data with at least one in depth (design, representation, execution or measurement)
3 - identifies all four limitations of data with at least one in depth or identifies 3 limitations of data with at least 2 in depth.
4 - identifies all four limitations of data with at least 3 described in depth

Appendix H: evaluative feedback matrix for assessments and inquiry reports

Section	notes	P	Feedback				
Hypo thesis	Lab		none	Why do you think this will happen	What do you think will happen?		
Proc	Lab		none	Procedures did not test hypothesis	Incomplete procedures		
Hyp test	Asses		none	Plan would not test hypothesis	Incomplete plan		
Explan claim	Both		none	Claim unclear			
Explan evidenc	Both		none	Vague reference to data	Not enough evidence	Evidence doesn't support claim	
Explan - reason	Both		none	Antidotal reasoning	Reasoning does not represent science	Reasoning does not support claim	
Pattern	Assess		none	Pattern doesn't match data	Pattern confusing		
error	Both		none	Missing design error	Missing measure error	Missing represent error	Missing executio n error

The idea behind this matrix is to create a sheet that could be stapled to a lab report or assessment. If sections are proficient, that box is checked. If sections are not proficient, applicable are highlighted. To help model what proficiency looks like, students with proficient answers would be invited to share. If there are no volunteers, the teacher could share examples or discuss what some of these comments mean. The purpose is to start the discussion to allow the teacher to ask the questions that prompt proficient work. This is an initial draft and would have to be used multiple times to create something that is viable for the next stage of research.