A Novel Approach to Fostering *Next Generation Science Knowledge* in Middle School Students: Introducing Double-Blinded Reviews in Classroom Formative Assessments

by

Sania Zahra Zaidi

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy (Educational Studies) in the University of Michigan 2016

Doctoral Committee:

Professor Nancy B. Songer, Drexel University, Co-Chair
Assistant Professor Leah A. Bricker, Co-Chair
Professor Brian P. McCall
Professor Kevin F. Miller

Dedication

Dedicated to the memory of my *nani* (grandmother) Dr. Jafri Begum who was the first woman in my family to earn a doctorate.

Acknowledgments

I am forever grateful to the many people who have supported me in this work. First, I would like to offer my gratitude to the participating teachers and students without whom this dissertation would not exist. Thank you for letting me into your classrooms and for enthusiastically engaging in the research process.

I would also like to thank my dissertation committee for always having my best interest in mind and providing me with valuable feedback and guidance. I would like to thank my academic advisor Nancy Songer for her mentorship and support throughout my doctoral program. I owe my scholarship to your mentorship. I am deeply indebted to you for the unwavering support you have provided me over the years. Thank you for the opportunities you have provided me to expand my scholarship and for always finding time for me even when you moved away to Philadelphia. This dissertation would not have been possible without your help, and for that I thank you.

I would like to thank Leah Bricker for being my mentor, for being by my side through this process and above all for being my friend. Thank you for spending countless hours with me in formulating my ideas and providing me with detailed feedback, especially on early drafts. I would also like to thank you for reminding me about the "light at the end of the (dissertation) tunnel", for all your encouragement and for sharing with me your own life experiences to help me on my journey. For this and much more, I am forever grateful to you.

I would like to thank Brian McCall for teaching me everything I know about quantitative research. Thank you for always finding time to meet me, for providing me with timely feedback even when you were on a sabbatical. Thank you for overseeing the numerous revisions I made to my quantitative models. I would like to thank Kevin Miller for offering me thoughtful and careful feedback. After taking your class in my second year, I knew I would gain immensely by having you on my committee. Thank you for being part of my dissertation committee and for supporting my scholarship. I would also like to thank Christina Weiland for being immensely understanding and giving me time off from the research project to work on my dissertation.

I would like to take this opportunity to thank my friends and family. To my parents, I am thankful for their constant love and encouragement. Thank you for instilling in me the value of education and for providing me with the opportunities that brought me to the University of Michigan. To my sisters, I am thankful for believing in me and for celebrating my accomplishments. To my friends, I am thankful for *always* being there for me and for putting up with me. Thank you for your love, patience, support and humor throughout this process. I could not have done this without you all.

Lastly, I would like to thank Rackham Graduate School for recognizing my scholarship and awarding me the Rackham Pre-doctoral fellowship. Thank you for equipping me with the means to complete my dissertation in a timely manner.

Table of Contents

List of Tables

Table

List of Figures

Figure

List of Appendices

Appendix

Abstract

This dissertation responds to the need in K-12 science education of developing classroom formative assessments and researching their role in promoting learning of integrated science knowledge. An innovative classroom formative assessment called the *4R Activity* is designed and implemented in middle school classrooms across the U.S.A. The 4R Activity was embedded in a climate change curriculum designed to foster integrated learning. In the *4R (Respond, Revise, Review, Reflect) Activity,* students engaged in the process of receiving and providing feedback, an activity that is central to the work of scientists and engineers during the construction of knowledge. The 4R Activity allowed students to receive peer and expert feedback on their knowledge products, which students then used to revise and resubmit their product. In an effort to promote receiving of unbiased feedback, the 4R Activity simulated the double-blinded review process that is widespread in academia. This dissertation presents three studies with separate but complementary areas of inquiry on the 4R Activity.

The first study investigated the efficacy of the 4R Activity in promoting integrated learning for *all* students. A quasi-experimental research design was employed to compare the achievement outcomes of 399 students distributed across the control and intervention groups. The intervention group comprised students (N=173) who used the 4R Activity while the control group students (N=256) used a conventional formative assessment within the same curriculum. A fixed-effects regression model was estimated

to evaluate the achievement outcomes of the two groups. Results revealed that all students in the intervention group outperformed (M=2.29; SD=1.99) the students in the control group. Furthermore, the 4R Activity had an additional positive effect on achievement gains of traditionally underrepresented groups in science and engineering (e.g., females, non-white racial/ethnic groups)

The second study documented the emerging patterns and characteristics of the work products generated by students who participated in the 4R Activity (N=173). Qualitative techniques were used to generate the patterns and characteristics. The findings revealed that over time students were able to use feedback to construct high quality integrated knowledge products. The reviews provided by students became increasingly elaborate and specific. Students began to articulate the aspects in the review that needed improvement as well as provided specific suggestions for strengthening the knowledge product. Similar to the progression in reviewing, students' reflections on their peer reviewing skills also became more elaborate and metacognitive. The reflections began to include self-assessment of reviewing skills and documentation of aspects to improve upon in future reviewing activities.

The third study used interview data from 22 students to present students' perspectives on learning from engaging in the 4R Activity. Students' accounts in this chapter contribute to the arguments made in research for the need to leverage students' experiences and interests to inform and promote learning in the classroom. Furthermore, this study also showcases how students' accounts can inform design decisions and successive iterations of learning environments. In this study, students' accounts provided insight into the connections students drew between the different stages of the 4R Activity and similar activities undertaken by scientists and engineers. The interview data also helped to showcase the connections students drew between the 4R Activity and their interests.

Together the three studies discussed in this dissertation help in showcasing how a classroom formative assessment like the 4R Activity can be used to promote learning, engage students in the practices of professionals and elicit students' interests and experiences that can be used productively to aid in future learning.

Chapter 1

Introduction

Science education in the United States is undergoing a wave of reform. Drawing from current research on learning and teaching this reform movement responds to the challenges facing the United States today, such as: a diminishing workforce in the fields of science, technology, engineering and mathematics (STEM) (see National Research Council [NRC], 2007); and underrepresentation of females and other demographic groups (e.g., non-white racial/ethnic groups) in postsecondary STEM education and careers (Burke & Mattis, 2007; NRC 2011). In an effort to address these challenges researchers and educators have proposed a dramatic shift in K-12 science education, which is outlined in the document, *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas* (National Research Council [NRC], 2012, hereafter referred to as the "*Framework*"). The *Framework* defines the overarching vision for K-12 science as follows:

> *"[To] ensure that by the end of 12th grade, all students have some appreciation of the beauty and wonder of science; possess sufficient knowledge of science and engineering to engage in public discussions on related issues; are careful consumers of scientific and technological information related to their everyday lives; are able to continue to learn about science outside school; and have the skills to enter careers of their choice, including (but not limited to) careers in science, engineering, and technology*." (NRC, 2012, p.1)

The vision of the *Framework* is translated into the *Next Generation Science Standards* (NGSS) (NGSS, Lead States, 2013). The NGSS emphasizes the building of science knowledge that is an integration of three dimensions: (a) science and engineering practices (e.g., constructing explanations and designing solutions), (b) disciplinary core ideas (e.g., core concepts in life science, physical science and earth science) and (c) crosscutting concepts (e.g., cause and effect relationships, patterns) (NGSS Lead States, 2013). Furthermore the NGSS advocates for fostering integrated knowledge, where disciplinary core ideas and crosscutting concepts are learned *through* direct engagement in the practices of science and engineering (NGSS Lead States, 2013;NRC, 2014; NRC, 2012). This integrated knowledge is also called, *Next Generation Science Knowledge* (NGSK) by Songer and colleagues (Songer, Zaidi & Newstadt, in preparation).

Fostering NGSK requires a rethinking of all aspects of science education such as curriculum, instruction and assessment (NRC, 2014). The role of assessments in building this new type of knowledge becomes pivotal because assessments can provide evidence of what point a student has reached in the development of NGSK. This information can be instrumental in further supporting students in deepening their NGSK within and across grades. In addition to developing NGSS aligned assessment tasks and accompanying scoring rubrics that provide evidence for students' development of NGSK, new designs of assessment activities that periodically gather this evidence also need to be developed (NRC, 2014). The design of new assessments to gather such evidence would require a shift in emphasis from large-scale assessment systems used for summative evaluations to classroom assessments that provide periodic evidence for students development of NGSK (see NRC, 2014). To support this shift the National Research Council (NRC, 2014) proposed a new path for assessments, "that is "bottom up" rather than "top down": one that begins with the process of designing assessments for the classroom, perhaps integrated into instructional units, and moves toward assessments for monitoring (NRC, 2014, p.8)

In the “bottom up” (NRC, 2014) approach to assessment design, classroom formative assessments can be pivotal in providing evidence for students’ current proficiency on NGSK while also aiding in informing further learning of NGSK. This is because *formative assessments* are carried out at intermediate time points during instruction to help make student thinking visible and provide feedback to help inform learning and teaching (Black & Wiliam, 1998). The feedback allows learners to become aware of the gaps that exist between the desired goal and their current understanding, and guides them to achieve the desired goal (Black & Wiliam, 1998; Darling-Hammond, &Bransford, 2007). Recognizing the important role that formative assessments can play in supporting learning of NGSK, the National Research Council (NRC, 2014) advocates for the need to design formative assessments to measure and support NGSK learning (NRC, 2014).

Responding to the need to design and research classroom formative assessments in K-12 science classrooms this dissertation presents three studies on a designed formative classroom assessment called the *4R Activity* (see next section, this chapter). Furthermore, the National Research Council (NRC, 2014) proposes that assessments be integrated into instructional units. The 4R Activity was integrated into an NGSS aligned middle school curriculum unit titled, *Climate Change and Impacts on Ecosystems* (Songer et al., 2014). The 4R Activity provides a model for a way science educators can begin to respond to the recommendations in NRC’s assessment report titled, *Developing Assessments for the Next Generation Science Standards* (NRC, 2014; hereafter referred to as the “assessment report” in the dissertation).

The 4R Activity

As a method to provide students with customized feedback that catered to their learning needs, the 4R Activity simulated the double-blinded review process widely used in academic journals. In addition to providing customized feedback, this feedback mechanism allowed students to directly participate in an activity that reflected a way in which scientists and engineers engage in building and communication of knowledge.

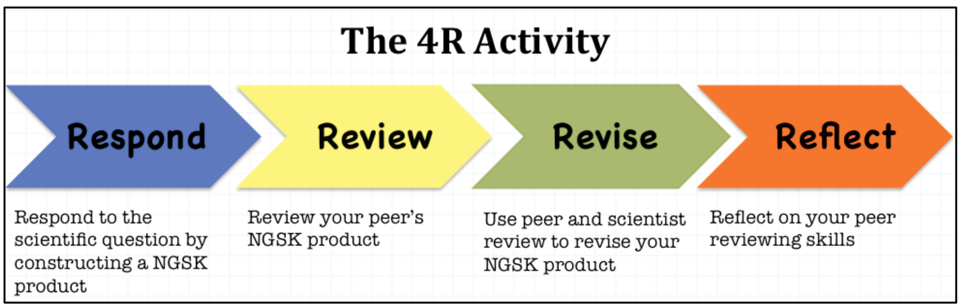


Figure 1.1. The four stages of the 4R Activity

The 4R Activity was designed and progressively refined through testing with professional scientists and piloting in one classroom over the course of one year (2013-2014). Pilot testing with six professional scientists from various disciplines (e.g., Chemistry, Biology, Neuroscience) was undertaken to gather their feedback on how to make improvements to the 4R Activity so that it reflected the peer review process of academic journals. The most salient feedback from the scientists was that there was a need to have at least two reviewers during the review process. This feedback was incorporated into the design process and the author of this dissertation took on the role of an additional reviewer, who was referred to as the "scientist" in the 4R Activity. Pilot testing in one classroom allowed to study the usability aspects of the 4R Activity and provided information on how to make the directions for the different stages of the 4R Activity clear to the students. The feedback from the pilot tests was then used to refine the 4R Activity. The 4R Activity in its present version that was used with the participants of this dissertation is described below.

The activity in its present version was carried out online over a two-day period. The students accessed the different stages of the 4R Activity by entering a code provided by the teacher prior to each stage of the 4R Activity. Students began the first stage of the 4R Activity by entering a computer generated four-digit code to access the activity prompt. In the first stage of the 4R Activity, students responded to the activity prompt by constructing their NGSK product. The author of this dissertation tracked the submission progress of the students. Once all students had submitted their NGSK products, students began with the review stage. In the review stage, students reviewed a peer's NGSK product corresponding to the same prompt. The author of this dissertation along with the teacher tracked the submission progress of the review. The review stage was completed when all students had submitted their review. After the students had completed the second stage, the author of this dissertation provided an additional review on each student's NGSK product. This review was referred to as the "scientist's review" in the 4R Activity. The scientist's review provided guidelines for revision based on the peer review and additional aspects that were not present in the peer review.

Day 2 of the activity began with the third stage of the 4R Activity. Students used the peer and scientist's reviews to revise the NGSK product in the third stage of the 4R Activity. The participating teacher in each classroom and the author of this dissertation tracked re-submission progress for the students. The end of the revise stage was followed by the reflection stage. In the reflection stage, students reflected on their reviewing skills by doing the following: (a) re-reading the review that they provided to their peer; (b) reading the scientist’s feedback on their peer reviewing skills; (c) and reading the review that the scientist gave to their peer. The reflection activity was designed to provide students access to an expert’s (i.e., the scientist) review on the same product that they themselves had reviewed.

The 4R Activity was a formative assessment that also functioned as a collaborative learning tool where collaboration with a peer and an expert could be facilitated to foster knowledge building. As previously noted, the 4R Activity was a translation of the review process undertaken by scientists and engineers. The review process in the 4R Activity differed from the double-blinded review of academic journals in the aspect that the submissions were not blinded from the “scientist” (author of this dissertation), who functioned as an additional reviewer. The scientist took on the role of a reviewer in addition to that of an editor because it was not feasible for students to construct an NGSK product and provide two reviews in one class period without compromising on the quality of the NGSK product or the review. However, as previously noted, discussions with professional scientists had made it apparent that it was important that students receive at least two reviews for it to be considered as a translation of the activity of peer reviews undertaken by scientists.

Addressing the Vision of Science Education through the 4R Activity

The *Framework* (NRC, 2012) draws from current research on learning to address the various weaknesses identified in K-12 science education, such as: (1) the lack od systematic organization of science learning across grade bands, (2) emphasis on discrete factual information focused on covering breadth over depth of concepts, and (3) lack of engaging opportunities provided that expose students to how science and engineering professionals engage in knowledge building and communication (NRC, 2012). Of the many salient themes emerging from the *Framework* to address these weaknesses, three overarching themes provide insight into how the vision of the *Framework* can be achieved.

First, the *Framework* emphasizes that science education should support *all* students in continually building on integrated knowledge (referred to as "NGSK" hereafter) (NRC, 2014; NRC, 2012). Furthermore the *Framework* emphasizes that that to build proficiency on NGSK would require repeated exposures over multiple years of schooling (NRC, 2012). Therefore the *Framework* has proposed a reduced set of disciplinary ideas in science that are re-introduced with increasing sophistication from kindergarten through twelfth grade (NRC, 2012). It is through this iterative process of knowledge building that the *Framework* intends for *all* students to develop a " more scientifically based and coherent view of the sciences and engineering" (NRC, 2012; p.11).

The second theme in the *Framework* is to engage students in real world activities undertaken by the scientific and engineering communities when building knowledge (NRC, 2012). The *Framework* emphasizes that learning experiences designed for students should engage them with fundamental questions about the world and expose them to the activities that scientists and engineers have engaged in to answer those questions (NRC, 2012). In other words, the learning environment should expose students to how knowledge is constructed in the scientific discipline and how scientific knowledge construction is a collaborative process, which is open to revision when new innovations and new theories are developed and new evidence emerges (NRC, 2012).

The third theme discussed in the *Framework* is to make science relevant to the lives of students so that learning of science is linked to learners' interests, experiences and aspired career choices (NRC, 2012). The *Framework* anticipates that making science relevant to the students' lives can aid *all* students in seeing the utility of science and engineering in tackling world challenges, can promote student interest in taking up science and engineering careers, and can equip students with sufficient knowledge to engage in science-related discussions and be critical consumers of scientific information related to their everyday lives (NRC, 2012).

This dissertation contends that the 4R Activity was a tool that science educators could use to begin to address the themes outlined in the *Framework* (NRC, 2012). Providing evidence of students' proficiency on NGSK as well as supporting students in deepening NGSK was the aim of the *4R Activity* (see Figure 1.1). Students NGSK was measured through the design of NGSK items and accompanying scoring rubrics that had been previously validated (see Songer, Zaidi, Newstadt, in preparation). As a process of providing feedback on the constructed NGSK, the 4R Activity employed the formal review system of academic journals. Employing this feedback mechanism allowed students to directly participate in an activity that reflected a way in which scientists and engineers engage in building and communication of knowledge. The results on students NGSK products and their reviewing skills are reported in this dissertation. Moreover this dissertation also reports on the effort toward making science learning relevant to the lives of the students. It presents an important first step of interviewing students to elicit their perspectives on the relevance of the activity to their lives and the connections they see between the 4R Activity and the work of scientists and engineers. Documenting students' accounts also marks an important first step in contributing the argument made by some researchers (e.g., Warren, Ballenger,Ogonowski, Rosebery& Hudicourt-Barnes, 2001; Nasir, Rosebery, Warren & Lee, 2006) on leveraging students' interests and experiences to promote learning in classrooms.

Learning Theories Guiding the Design of the 4R Activity

The design of the 4R Activity was guided by the learning theories of social constructivism and situated cognition.

Social Constructivism

The 4R Activity was designed under the social constructivism perspective that acknowledges both Piagetian and Vygotskian influences. Social constructivism emphasizes the interdependence of individual cognitive processes (Piaget, 1970) and social processes (Vygotsky, 1978) in the construction of knowledge (see Linn & Eylon, 2006). According to the social constructivist perspective, the social context in which learning occurs is regarded as critical, as learning is contextually specific (e.g., Palincsar, 1998). As learners participate in collaborative activities they internalize the strategies and knowledge, which they can then use in future independent problem-solving activities (Vygotsky, 1978; Palincsar, 1998). Additionally, learning in the social context is mediated through the use of tools (physical and psychological) and is key to the construction of knowledge (Vygotsky, 1978; 1981). Furthermore, the use of language as a tool has a central role in social constructivism and is often considered as one of the most important tools (Vygotsky, 1978; Rogoff, 1990; Wertsch, 1994). Language as a tool not only serves in learning the use of other tools but it is also vital in how learners express their thinking and internalize knowledge and tool use (Rogoff, 1990; Karpov, 2003).

The 4R Activity was designed with the social constructivist perspective in mind. In the 4R Activity learning was seen as occurring during the collaborative activity of receiving and providing critique. Written discourse was used as learners received and provided critique and co-constructed knowledge. Furthermore the use of technology mediated how learners interacted with each other (e.g., anonymously) and could also be considered a tool that mediated student learning. Additionally, the 4R Activity as a whole could also be viewed as a tool that mediated learning of NGSK while also helping students internalize tool use, i.e., the activities in different stages of the 4R Activity (e.g., review, reflection, revision).

Situated Learning

The 4R Activity is grounded in the perspective of situated learning (Brown, Collins & Duguid, 1989). Following the work of innovative thinkers such as Vygotsky and Dewey, Brown and colleagues (1989) argued that meaningful learning takes place when it is embedded in the social context within which it will be used. As a means of creating authentic activities, they proposed a model based on traditional apprenticeship, called *cognitive apprenticeship.* Cognitive apprenticeship replicates the critical elements of actual apprenticeship for the learner by creating a meaningful social context in which learners are given many opportunities to observe and learn the practices of the community. Through enculturation, learners adopt the norms, behaviors, skills, language and beliefs of a particular community (Lave & Wenger, 1991; Rogoff, 1990). Learning through this process allows individuals to “acquire, develop, and use cognitive tools in authentic domain activity” (Brown, et al., 1989, p.39). In the 4R Activity, the peer review system simulated the critical aspects of the review system of scientific journals, wherein learners submitted their knowledge product, received blinded reviews and guidelines for revision, and were provided an opportunity to use the feedback from the reviews to revise and resubmit their product.

One feature central to situated learning is that learning should support the collaborative construction of knowledge in authentic contexts that allows individuals to reflect and articulate their knowledge (Collins, Brown &Newman, 1989; Bransford, Vye, Kinzer & Risko, 1990). In situated learning, the activities are typically situated in complex social environments (Anderson, Reder & Simon, 1996), however access to expert performance, coaching and scaffolding are provided to the learners (Collins, et al., 1989; Griffin, 1995). Access to expert performance was provided in the 4R Activity by allowing learners to see the scientist’s review that their peer received. This allowed learners to compare and contrast the scientist’s review with the review that they had provided. The reflection activity on improving the peer review skills allowed learners to reflect on the aspects they wanted to improve upon and articulate the strategies that they would employ to improve their peer reviewing skills. Additionally scaffolding was provided to the learners throughout the activity, such as providing templates for constructing NGSK products (e.g., template that outlined parts of the NGSK product: claim, reasoning and evidence) and, providing them scaffolds during peer review (see Appendix C).

Increasingly researchers (e.g., Bransford, et al., 1990; McLellan, 1991) agree that computer-based representations provide a powerful and acceptable vehicle for situated learning in the classroom environment. For example, Reeves (1993) concludes that a major benefit of an interactive multimedia environment is its ability to include opportunities for simulated apprenticeships together with a wealth of activities supporting learning. Building on this argument, the 4R Activity was designed as a computer-based representation, which facilitated in blinding the reviews, tracking student progress during the activity and providing students reviews in a timely manner.

Implementation Details of the 4R Activity

This section presents the details pertaining to the administration of the 4R Activity such as details on the curriculum used for the 4R Activity, the time point of administering the 4R Activity as well as the activity prompts used in each stage of the 4R Activity.

The Curriculum

The 4R Activity was embedded in the *Climate Change and Impacts on Ecosystems* (Songer et al., 2014) middle school online curriculum. The curriculum was aligned to the NGSS (NGSS Lead States, 2013) and was designed to promote learning of NGSK (Songer, Zaidi, Newstadt, in preparation). In the curriculum, the NGSK product was introduced to the students as an explanation or a prediction, where predictions comprised events that occurred in the future such as the affects of climate change on species distribution in the future. Both the explanations and the predictions were written in a format that comprised a *claim* supported by relevant *evidence* and scientific *reasoning* that linked the claim to the evidence. This format was similar to the format of explanation used by the other researchers in science education (e.g., Songer, Kelcy & Gotwals, 2009; Songer & Gotwals, 2012; McNeill, Lizotte, Krajcik & Marx, 2006).

Administering the 4R Activity

Students participated in the 4R Activity after engaging with one of the three focal questions in the curriculum. The lessons in the curriculum were organized under three overarching question (referred to as focal questions). The three focal questions were: (1) *How do biotic and abiotic factors affect where species live?;* (2) *How has Earth's temperature changed in the last 150 years?; and* (3) *How will climate change affect species distribution?* The lessons within each focal question were designed to promote learning of NGSK. Students constructed multiple NGSK products as they progressed through the lessons. In the 4R Activity students constructed an NGSK product pertaining to the focal question that had been addressed by the lessons preceding the 4R Activity. Figure 1.2 provides an example of a focal question, the activities in the curriculum addressing that focal question, along with the 4R Activity that students engaged with after the focal question was addressed in the curriculum.

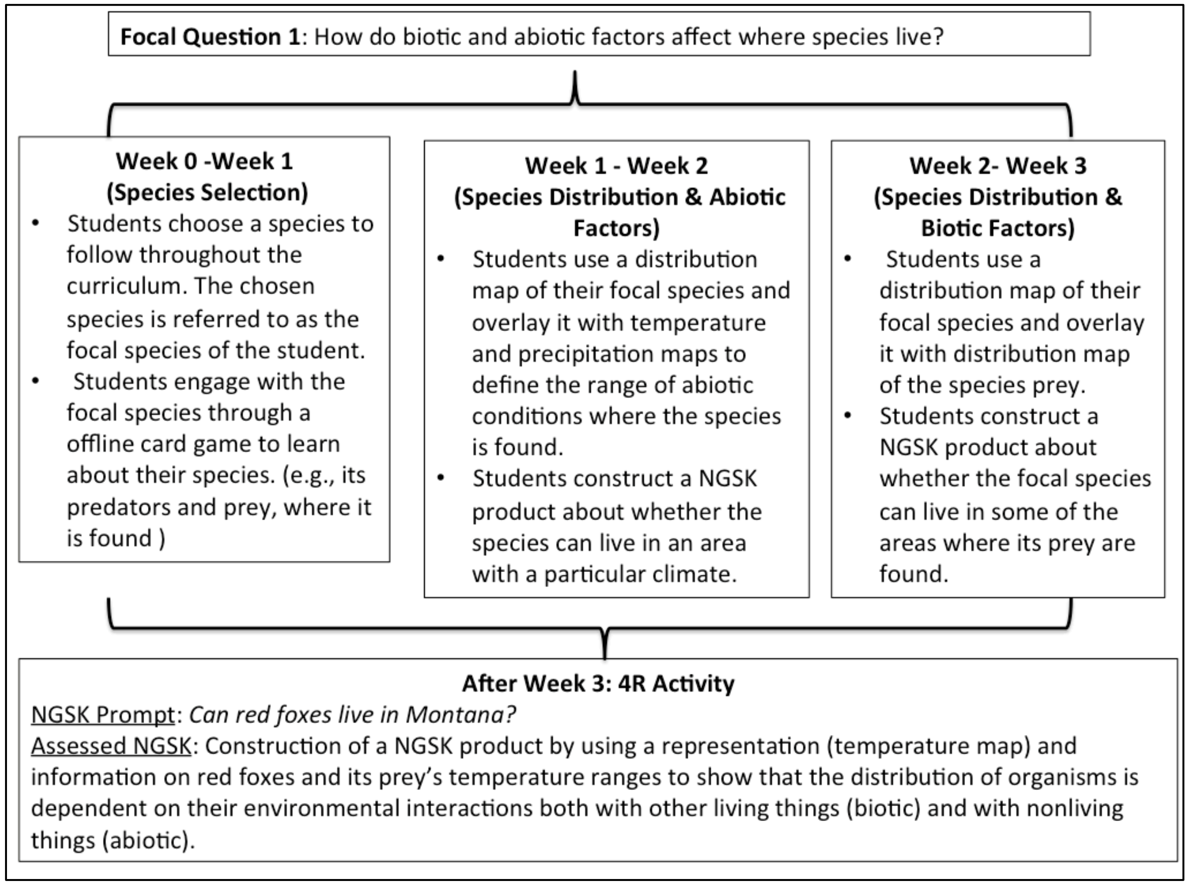


Figure 1.2. An example of a focal question, the activities in the curriculum addressing the focal question, and the 4R Activity administered after the focal question.

In the week prior to the first 4R Activity, an offline practice activity was set up to introduce the peer review sheet (refer to Appendix C) used in the 4R Activity. In the practice activity, students worked in groups of three to four to review an already constructed NGSK product. The NGSK product provided for the review activity pertained to a scientific question that the students had previously answered in the lesson preceding the review activity. After the student groups had reviewed the NGSK product, they shared their reviews with the class and discussed suggestions for improvements, such as suggestions pertaining to the use of specific evidence, reasoning etc. By the end of the practice activity students had worked with the peer review sheet and has participated in a classroom discussion. The discussion provided opportunities for students to learn to use the different areas of improvement listed in the peer review sheet, as well as some examples of how to provide suggestions for a particular area of improvement.

After the practice activity students continued with the curriculum for another week before the first 4R Activity was administered. As previously discussed, the 4R Activity was administered over a two-day period. Figure 1.3 represents the administration timeline of the 4R Activity. In Figure 1.3, the left hand side represents the timeline for the administration of the 4R Activity within the curriculum while the right hand side represents the timeline for the different stages of the 4R Activity carried over the two-day period.

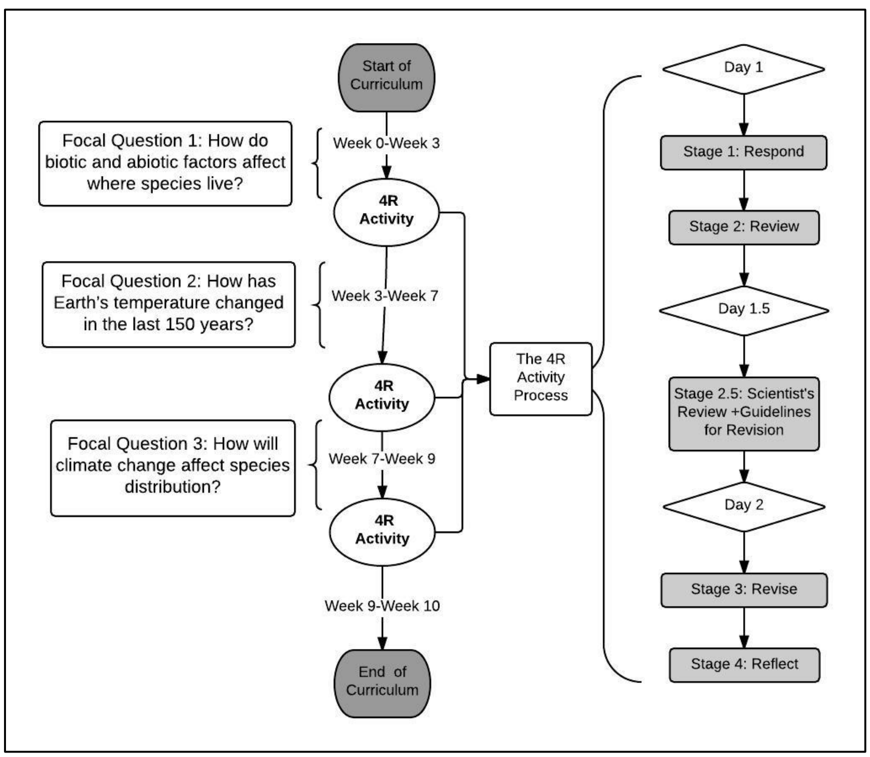


*Figure 1.3.*The implementation of the 4R Activity within the curriculum.

Description of the Activity Prompts used in the 4R Activity

As shown on the right-hand side of Figure 1.3, there were four stages in the 4R Activity. Each R in the 4R Activity indicated one of the four stages in the activity: Respond, Revise, Review, and Reflect. A detailed description of the prompts in each stage of the 4R Activity is presented below.

Stage 1: Respond. In the first stage of the 4R Activity students constructed an NGSK product in relation to a focal question. The activity prompts in the 4R Activities for constructing the NGSK products are shown in Table 1.1.

Table 1.1

Activity prompts for constructing the NGSK products in the 4R Activity

Focal Question	**4R Activity**	**Activity prompt for the NGSK product**	**Assessed NGSK**
1. How do biotic and abiotic factors affect where species lives?	Activity 1	*1. Can red foxes live in Montana?*	Construction of a NGSK product by using a representation (temperature map) and information on red foxes and its prey's temperature ranges to show that the distribution of organisms is dependent on their environmental interactions both with other living things (biotic) and with nonliving things (abiotic).
2. How has Earth's temperature changed in the last 150 years?	Activity 2	*2. If carbon dioxide emissions continue to increase in the future, what do you predict will happen to the Earth's surface temperature?*	Construction of a NGSK product by analyzing and interpreting temperature data of the past century to predict the effect of increasing carbon dioxide emissions on Earth's mean surface temperature.
3. How will climate change affect species distribution?	Activity 3	*3. Will brown squirrels be found in Ohio in the future?*	Construction of a NGSK product by using graphical displays of large datasets as evidence to show the cause and effect relationship between organisms and their interaction with nonliving factors.

When responding to each activity prompt, a student had access to artifacts such as maps and graphs to better support him/her in construction of the NGSK product. To provide more information on the difficulty of the NGSK being assessed in each of the three 4R Activity prompts, item response theory (Hambleton, Swaminathan, & Rogers, 1991) was employed by using student data from a previous implementation cycle. The analysis revealed that the most difficult NGSK product was for activity 2, which was three times as difficult as the NGSK product constructed in activity 1. Also, the NGSK product constructed in activity 3 was twice as difficult as that of activity 1 (see Chapter 2, in this dissertation). The information was used to do comparison analysis of students' relative scores on the NGSK products. Thus the improvement in the scores could also be interpreted in light of the difficulty of the items.

Stage 2: Review. The review process was undertaken in the second stage of the 4R Activity. In this stage students engaged in a double-blinded peer review. In an effort to support students in learning to provide peer reviews a pre-designed review sheet was provided to each student (see Appendix C). The review sheet simplified the review process by listing five areas of improvement that peers could make along with a suggestion template pertaining to each area. The five areas of improvement were:

A: Not all parts of the explanation/prediction are present

B: Explanation/prediction does not have specific evidence

C: Explanation/prediction does not have specific reasoning

D: Explanation/prediction needs stronger scientific words

E: Other Improvement

The “Other Improvement” category typically encapsulated grammatical errors, syntax, spelling errors and suggestions pertaining to improvement in sentence construction. Prior to the 4R Activity, each teacher and his/her students discussed some aspects to consider for this category, such as capitalization of words, grammar, punctuation and spelling errors. Overall across the five classrooms, the “Other Improvement” category primarily functioned as a category for improvements related to writing.

From the five areas of improvement, the reviewers were required to select one area of improvement on which they wished to focus their reviews. This was referred to as the primary area of improvement in the peer review. Reviewers were also encouraged to select an additional area of improvement but the selection of an additional area was optional.

Stage 2.5: The Scientist’s Reviews. As previously discussed in the description of the 4R Activity, a scientist (the role undertaken by the author of this dissertation) provided an additional review to each student to complement the peer review. The review provided by the scientist was done in a format similar to that of an editor’s review in academic journals. The scientist’s review listed the suggested steps for revision for the author. In the suggested steps, the review given by the peer reviewer was emphasized and additional suggestions were provided (if necessary). In addition to providing feedback to the author, the scientist also provided brief feedback to the reviewer that listed aspects that the reviewer could work upon in future peer reviewing.

Stage 3: Revise. The revision activity was carried out on day 2 of the 4R Activity. During revision of the NGSK products, students had access to the reviews that they received from a peer and the scientist. Students had access to their original NGSK product and they could make whatever revisions they considered necessary based on the reviews they received. Re-submission of their revised NGSK product marked the end of the revision activity.

Stage 4: Reflection. Reflection was designed to provide support to students' reviewing skills. Figure 1.4 shows the reflection activity prompt that the students received in the 4R Activity.

Reflection

1. Re-read the explanation that you reviewed.
2. Re-read the review that you gave to your peer.
3. Now read the scientist's feedback on your review and the scientist's review to your peer.
4. Write 1 or 2 things that you learned about improving your peer reviewing skills from reading the scientist's feedback and review.

Figure 1.4. Instructions provided to students in the first reflection activity.

In the reflection activity, each student read the feedback they received from the scientist on his/her reviewing skills. Additionally, every student also read the scientist's review on his or her peer's NGSK product. The reflection activity began with students re-reading the NGSK product that they reviewed along with their review of that product. This was followed by them reading the feedback on reviewing skills along with the scientist's review of that same NGSK product. Students were then prompted to reflect on what they had learned from reading the scientist's feedback to them about their reviewing skills and the scientist's review of the NGSK product that they both reviewed.

Overview of Chapters

The forthcoming chapters in the dissertation report on three aspects of the 4R Activity that were investigated. Chapter 2 is a report on achievement outcomes of students who participated in the 4R Activity. Next, Chapter 3 is an in-depth examination of the emerging patterns and the changing characteristics of students' artifacts (e.g., revised NGSK products, peer review and reflection) over the course of three 4R Activities. Chapter 4 reports on students' perspectives on the 4R Activity in relation to the 4R Activity advancing their learning, its relevancy to their lives and its connections to the work of scientists, engineers and other professionals. The dissertation concludes with Chapter 5 where major contributions, next steps for research, and concluding remarks are discussed.

Chapter 2

The focus of Chapter 2 is to discuss the efficacy of the 4R Activity in promoting proficiency on NGSK for *all* students. This chapter presents the results of a quasi-experimental study on the achievement gains of two groups of students (N=399) that participated in the CCIE (Songer et al., 2014) curriculum. The "treatment" administered to the two groups was the difference in the feedback process (i.e., providing and acting on feedback) during the formative assessment intervention. The group using the 4R Activity was provided with individualized feedback through the aforementioned review process, which the students then used to revise and resubmit the NGSK products. In contrast, the other group used the 1R Activity where providing and acting on feedback took the form of a teacher facilitated classroom discussion on observed trends in students' written NGSK products along with strategies on improving the quality of the NGSK product.

A fixed effects model was used to determine the achievement gains for different sub-groups of the population, such as female, racial-ethnic minorities, English language learners (ELLs), and students with Individualized Education Programs (IEPs). This allowed for determining the variability in the effects of group assignment on these attributes. Additionally the difference in achievement gains for each of the three 4R Activities was also analyzed separately. This allowed for determining the variability in achievement based on the difficulty of the items administered during the three 4R Activities. This chapter discusses the implication of these findings in light of promoting learning for all students while also examining the achievement outcomes of students who have been traditionally underrepresented in STEM fields.

Chapter 3

Chapter 3 presents a study to showcase how a formative assessment system like the 4R Activity can aid in promoting NGSK learning through the use of disciplinary practices, such as receiving and providing critique. This chapter presents students' data (N=173) from the three 4R Activities embedded throughout the CCIE curriculum. It discusses the changing patterns and characteristics in students' revised NGSK products, peer reviews and reflections on their own peer reviewing skills. The patterns and characteristics were compared across the three 4R Activities for students in classrooms of the five participating teachers. The patterns and characteristics were derived using qualitative analysis (e.g., Maxwell, 2013; Patton, 2002), where student data was first open coded and then focused codes were developed using established qualitative techniques. The chapter discusses the implications of these findings in light of how classroom assessments like the 4R Activity can engage students with the work of scientists and engineers while also fostering the development of proficiency on integrated knowledge like NGSK.

Chapter 4

This chapter uses interview data of twenty-two students to discuss their perspectives on learning from engaging in the 4R Activity, as well as the applicability they see of the 4R Activity relative to their interests (e.g., playing basketball). The *Framework* advocates for the use of classroom assessments to elicit students' interests and experiences so that they may be effectively used to meet the goals of instruction (NRC, 2014). However the use of classroom assessments to elicit students' experiences and interests has not been a focus in assessment research (NRC, 2014). This study responds to this gap in classroom research by providing an example of a formative assessment that students used to reflect on applications to their out-of-school activities, participation in other school subjects, and the like. Furthermore, student accounts provide insight into the connections students draw between the 4R Activity and the activities undertaken by scientists and engineers. Qualitative techniques (e.g., Maxwell, 2013) were employed to draw themes emerging from student accounts. This chapter discusses how assessments can contribute towards eliciting students' interests and experiences and aids in providing valuable feedback to the teacher for forthcoming instructional activities that make use of students' prior experiences and interests.

Chapter 5

The sum of the above three chapters makes a case for the use of a classroom assessment system like the 4R Activity to assess and foster Next Generation Science Knowledge (NGSK) for all students. The final chapter discusses, the contributions of this dissertation to the research community, the limitations of the studies and the next steps for the 4R Activity. The chapter concludes with offering implications for science educators and formative assessment research.

References

Anderson, J. R., Reder, L. M., & Simon, H. A. (1996). Situated learning and education. *Educational researcher*, *25*(4), 5–11.

Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher*, *18*(1), 32–42.

Burke, R.J., and Mattis, M.C. (Eds.). (2007). *Woman and minorities in science technology, engineering, and mathematics: Upping the numbers.* Northampton, MA: Edward Elgar.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education:Principles, Policy and Practice, 5*(1), 7–73.

Bransford, J.D.,Vye, N., Kinzer, C., & Risko, V. (1990). Teaching thinking and content knowledge: toward an integrated approach. *Dimensions of thinking and cognitive instruction*. (pp. 381–413). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Collins, A., Brown, J. S., & Newman, S. E. (1989). Cognitive apprenticeship: Teaching the craft of reading, writing, and mathematics. In L. B. Resnick (Ed.), *Cognition and instruction: Issues and agendas* (pp. 453–494). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Darling-Hammond, L., & Bransford, J. (2007). *Preparing teachers for a changing world: What teachers should learn and be able to do*. John Wiley & Sons.

Griffin, M. M. (1995). You can't get there from here: Situated learning, transfer and map skills. *Contemporary Educational Psychology, 20,* 65–87.

Hambleton, R., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory*. London: Sage.

Karpov, Y. (2003). Vygotsky's doctrine of scientific concepts: Its role for contemporary education. In A. Kozulin, B. Gindis, V. Ageyev & S. Miller (Eds.), *Vygotsky's educational theory in cultural context* (pp. 138-155). Cambridge: Cambridge University Press.

Linn, M.C. & Eylon, B.S. (2006). Science Education: Integrating Views of Learning and Instruction. In P. A. Alexander & P. H. Winne (Eds.), Handbook of Educational Psychology (2nd Ed., pp. 511-544). Mahwah, NJ: Lawrence Erlbaum Associates.

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

Maxwell (2013). *Qualitative research design: An interactive approach.* Thousand Oaks, CA: Sage.

McLellan, H. (1991). Virtual environments and situated learning. *Multimedia Review, 2(3),* 30–37.

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

Nasir, N. I. S., Rosebery, A. S., Warren, B., & Lee, C. D. (2006). Learning as a cultural process: Achieving equity through diversity. *The Cambridge handbook of the learning sciences*, 489-504.

National Research Council. (2007). *Taking science to school: Learning and teaching science in grades K-8.* Committee on Science Learning, Kindergarten Through Eighth Grade. R.A. Duschl, H.A. Schweingruber, and A.W. Shouse (Eds.). Board on Science Education, Center for Education. Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

National Research Council. (2011). *Expanding underrepresented minority participation: America's science and technology talent at the crossroads.* Committee on Underrepresented Groups and the Expansion of the Science and Engineering Workforce Pipeline. Washington, DC: The National Academies Press.

National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas.* Committee on Conceptual Framework for the New K-12 Science Education Standards. Board on Science Education. Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

National Research Council. (2014). *Developing Assessments for the Next Generation Science Standards.* Committee on Developing Assessments of Science Proficiency in K-12. Board on Testing and Assessment and Board on Science Education, J.W. Pellegrino, M.R. Wilson, J.A. Koenig, and A.S. Beatty, *Editors.* Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

NGSS Lead States (2013). *Next Generation Science Standards: For States, By States.* Washington, DC: The National Academies Press.

Piaget, J. (1970). Piaget's Theory. In L. Carmichel (Ed.), *Carmichael's Manual of child psychology.* Chicago: John Wiley and Sons.

Patton, M.Q. (2002). Qualitative Research and Evaluation Methods (3rd ed.). Thousand Oaks, CA: Sage Publications.

Palincsar, A.S. (1998). Social constructivist perspectives on teaching and learning. *Annual Review of psychology*, *49* (1), 345–375.

Reeves, T. C. (1993). Evaluating interactive multimedia. In D. M. Gayeski (Ed.), *Multimedia for learning: Development, application, evaluation,* Englewood Cliffs, NJ: Educational Technology Publications, 97–112.

Rogoff, B. (1990). *Apprenticeship in thinking: Cognitive development in social context*. Oxford University Press.

Songer, N.B., Dewey, T., Peters, V., Reicher, M., Kwok, A., Zaidi, S., Tupper, B., Hammond, G., Myers, P. (2014) *Climate Change and Impacts on Ecosystems: A Middle School Curricular Unit.* Ann Arbor, MI: The University of Michigan.

Songer, N.B., Kelcey, B., & Gotwals, A. (2009). How and when does complex reasoning occur? empirically driven development of a learning progression focused on complex reasoning About Biodiversity. *Journal of Research in Science Teaching, 46*(6), 610 - 633.

Songer,N.B., & Gotwals, A.W. (2012). Guiding explanation construction by children at the entry points of learning progressions. *Journal of Research in Science Teaching, 49(2),* 141–165.

Songer, N.B., Zaidi, S., Newstadt, M. (2016). Classroom-based holistic assessment and evidence of middle school students' Next Generation Science Knowledge. Manuscript in preparation.

Vygotsky L.S. (1981). The instrumental method in psychology. In *The Concept of Activity in Soviet Psychology,* ed. J Wertsch, pp. 3–35. Armonk, NY: Sharpe.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard Univ. Press.

Warren, B., Ballenger, C., Ogonowski, M., Rosebery, A. S., & Hudicourt-Barnes, J. (2001). Rethinking diversity in learning science: The logic of everyday sense-making. *Journal of research in science teaching*, *38*(5), 529-552.

Wertsch, J.(1994). Mediated action in sociocultural studies. *Mind, Culture, and Activity, 1*, 202-208.

Chapter 2

Evaluating the Science Achievement of Middle School Students Who Used a Formative Assessment System Providing Customized Feedback and Revision Opportunities

Introduction

The assessment report titled, *Developing Assessments for the Next Generation Science Standards* (NRC, 2014, hereafter referred to as “the assessment report”) outlines the vision for science assessments to measure and support the learning of integrated knowledge like the NGSK. The assessment report emphasizes the role of classroom assessments in supporting NGSK. It calls for the development of formative assessments that provide periodic evidence of students’ NGSK. Furthermore it advocates for formative assessments that provide support to *all* students in deepening their NGSK, especially those student groups that are traditionally underrepresented in science and engineering careers (e.g. females, non-white racial and ethnic groups, English Language Learners, students with special needs) (NRC, 2014). Responding to the need to design classroom formative assessments that support all students in building NGSK, this study compares two types of formative assessments to showcase their effectiveness in stimulating achievement gains on NGSK for *all* middle school students.

Formative Assessments

Formative assessments employed in the classroom can serve the purpose of providing evidence of NGSK learning while also supporting students in building their NGSK. Research on formative assessment has shown the potential of formative assessments to aid in student learning (see Black & Wiliam, 1998a; Sadler, 1989; Bell & Cowie, 2001). For an assessment to be considered formative, first, the purpose of the assessment should be to inform teaching and/or learning. Second, formative assessment is rooted in the idea of feedback, which should be used to advance learning and/or teaching (Black &Wiliam, 1998b; Sadler, 1989). In other words, formative assessment serves the purpose of providing evidence of the gap between the current level of performance of a learner and the desired level of performance on a task while promoting learning through feedback directed on closing the performance gap (Black &Wiliam, 1998b; Shute, 2008).

Bell and Cowie (2001) described formative assessment to inform student learning as a cycle comprising three distinct activities: (1) gathering information on student learning; (2) interpreting the gathered information; (3) and acting on the gathered information with the purpose of improving student learning. Information can be gathered through the design of effective assessment tasks to assess the target knowledge while a rubric can be used to interpret student responses and determine the student's level with respect to the knowledge being assessed (Black & Wiliam 1998b). "Acting on the gathered information" can be achieved by providing feedback to the learner that is then used by the learner to inform his/her learning (Bell and Cowie, 2001;Black & Wiliam, 1998b). Moreover Black and Wiliam (1998a) found that it was "acting on the gathered information” that was a crucial step in supporting student learning and improving student achievement.

However, “acting on the gathered information” has been documented to be the most challenging step in the implementation of formative assessments (e.g., Furtak, Ruiz-Primo et al., 2008; Ruiz-Primo, Furtak, Ayala, Yin &Shavelson, 2010). Some aspects that contribute to the challenge are: (1) the effectiveness of a type of feedback provided (e.g., general vs. specific); (2) the effectiveness of the feedback for the diverse learning needs of learners; (3) the effectiveness of the feedback for the type of learning being fostered (Shute, 2008; Topping, 2010; McMillan, 2010; Narciss & Huth, 2004; Hattie & Timperley, 2007; Furtak, Ruiz-Primo et al., 2008); and (4) the effectiveness of the strategies employed to use feedback to support further learning (Hattie & Timperley, 2007; Nyquist, 2003).

Given the pivotal role of feedback (i.e., providing and acting on feedback) in supporting student learning it is imperative to conduct comparative studies of feedback systems (i.e., providing feedback and using feedback to inform learning) in formative assessments. Responding to the need of exploring the effectiveness of different feedback systems in promoting the learning of NGSK, this study compares two types of feedback systems, collective and individual. The formative assessments employing these feedback systems are written and embedded in a middle school climate change science curriculum (Songer et al., 2014). The curriculum employs the vision of the *Framework* (NRC, 2012) and promotes learning of NGSK in students.

In the first feedback system (called the 1R Activity), the process of providing and acting on the feedback is done at the classroom level through a teacher facilitated discussion, thereby making it collective in nature. The second type (called the 4R Activity) adopts an individual perspective, where customized feedback is provided to each student, which is then used by the student to revise and resubmit the NGSK products.

The following research question guided the comparative analysis of these two feedback systems:

To what extent do students participating in formative assessments using an individualized feedback system demonstrate achievement gains that are different from students that participate in formative assessments using a collective (e.g., classroom level) feedback system?

Relevant Background Research

As previously outlined, the formative nature of assessment hinges on the purpose of the assessment; that is assessments become formative when the information is gathered and used to advance teaching and learning (Black & Wiliam, 1998a; Sadler, 1989; Bell & Cowie, 2001; Popham, 2008). Research outlines multiple ways in which information can be gathered to inform teaching and learning. Shavelson and colleagues (Shavelson, 2005; Shavelson et al., 2008) placed the different types of formative assessment onto a continuum called the *formative assessment continuum* (Shavelson et al., 2008). On the left end of *formative assessment continuum* lie informal or unplanned formative assessments while formal or planned formative assessments lie at the right extremity of the continuum. Between these two extremities lie the planned-for-interaction formative assessments (Shavelson et al., 2008). The informal/unplanned assessments are mainly improvisational, where the teacher creates these assessments during instruction when he/she encounters that there is a need to elicit student thinking and learning (Shavelson et al., 2008). Like informal assessments, planned-for-interaction formative assessments are interactive assessments to elicit student thinking during instruction but unlike informal assessments, these are purposefully planned by the teacher in advance to elicit student thinking at specific time points in the instruction. Both informal and planned-for-interaction assessments are administered when immediate feedback is required by the teacher to assess in-the-moment student learning before moving forward with further instruction.

On the other side of the assessment continuum lie formal formative assessments. Similar to planned-for-interaction assessments these are also planned assessments that are administered at pre-defined time points in instruction to illicit evidence of student learning. Typically formal assessments are administered as curriculum-embedded assessments after a focal idea in the curriculum has been addressed and there is a need to assess student learning to ensure that students are ready to move forward in the curriculum. Thus a point of departure between formal and informal or planned-for-interaction assessments is that informal assessments might be administered whilst a certain idea *is being* introduced to students while formal formative assessments might be most useful *after* the focal idea has been introduced. Researchers have argued that formal assessments, such as the curriculum-embedded formative assessments allow teachers to more formally gather evidence of student understanding at critical junctures in the curriculum where student learning needs to be assessed before moving forward in the curriculum (Ruiz-Primo et al., 2010; Furtak & Ruiz-Primo, 2008). Given that the assessment is embedded into the curriculum, it allows teachers to make informed decisions on instructional adaptations based on formal evidence collected from all students (Ruiz-Primo et al., 2010).

However research on curriculum-embedded assessments remains sparse, especially in science with only a few notable exceptions (Furtak & Ruiz-Primo, 2008; Shavelson et al., 2008). This study argues that using curriculum-embedded formative assessments in science is important now more than ever given the relative newness of the Next Generation Science Standards (NGSS) (NGSS Lead States, 2013) and the need to both support and collect formal evidence of students learning of NGSK. Responding to this need this study explored two types of curriculum-embedded formative assessments in the middle school curriculum, *Climate Change and Impacts on Ecosystems* (Songer et al., 2014). Through the comparative analysis of the two assessment systems, this study discusses effective ways of implementing curriculum-embedded formative assessments in light of supporting NGSK learning of students.

Feedback Research and Formative Assessments

In a classroom there are students with diverse learning needs and therefore it is important to design environments that address the needs of learners and maximize their learning. Research on formative assessment has shown to improve performance of low-achieving students and students with identified learning difficulties (see Black &Wiliam, 1998b). Shepard (2008) explains this finding by stating that formative assessments explicitly provide resources to low-achieving students that fosters the development of metacognitive skills and increases motivation, which might not be available to them otherwise. Many of these effects of formative assessment are tied to the process of feedback, which lies at the heart of formative assessment (Wiliam, 2010).

Feedback can be seen as the most important tenet of formative assessment and it is has been extensively studied. Hattie and Jaeger (1998) define feedback as information provided to assist learners in meeting the specified learning goals of the activity. Different types of feedback have shown to be useful for different types of learning (Shute, 2008; McMillan, 2007; Wiggins &McTighe, 2005). For example, when assessing "simple knowledge" (e.g., recall of facts) (McMillan, 2007) the feedback provided is on correct versus incorrect information. On the other hand assessments for "deep understanding" (McMillan, 2007; Wiggins & McTighe, 2005) that require integrating multiple pieces of knowledge, problem solving and critical thinking need to focus on feedback that emphasizes the closing of the learning gap. Sadler (1998) suggests that feedback that emphasizes closing of this learning gap, can take the form of oral or written feedback that is accessible to the students. Such feedback makes apparent where the learning gap exists and provides concrete suggestions on how to close the learning gap. Feedback can also be provided by multiple sources and in multiple ways to narrow the learning gap. Research shows that students more readily incorporate individualized feedback (Shute, 2008; Narciss &Huth, 2004). Moreover, individualized feedback is shown to specifically help low-achieving students while higher achieving students may require less specific feedback (Mason &Bruning, 2001; Shute, 2008).

There is also a component within formative assessment of who gives feedback. Peer involvement in assessment is encouraged as it not only builds the evaluative skills of assessors, but it also presents feedback from a different perspective than a teacher or an expert would (Topping, 2010). This is often the case as peers might be operating from a similar frame of reference as the learner and might be successful in providing feedback that is more accessible to the learner (Topping, 2010; Brown, Rust &Gibbs, 1994). Within written assessments, peer feedback's effectiveness is substantial in research (Yang, Ko &Chung, 2005; O'Donnell &Topping, 1998; Richer, 1992; Ramsden, 2003). Additionally, peer reviewing has shown to benefit the reviewer's own writing (Lundstrom & Baker, 2009). However, one major disadvantage of peer review is the negative effects of peer review arising from the social dynamics of classrooms. These dynamics could result out of friendships or enmity between peers and/or due to gender and racial-ethnic stereotypes in a classroom (Topping, 2010; Schaffer, 1996; Topping, 1998). Researchers suggest the use of anonymous review system to alleviate the potential threat of negative effects of peer reviewing (Jessup, Connolly & Tansik, 1990). Anonymous reviewing has shown to encourage more constructive feedback from reviewers (MacLeod, 1999; Robinson, 1999; Lu & Bol, 2007).

As discussed above, feedback is regarded as crucial to improving learning (see Azevedo & Bernard, 1995; Moreno, 2005) and different features of feedback have shown to separately impact different types of learning. However there is little empirical evidence on the comparison of the different features of feedback as they interact with learner characteristics (e.g. learning needs of students, race, gender) and the instructional context (e.g., discipline-specific knowledge) (see Shute, 2008; Narciss & Huth, 2004). This study responds to the need for research on comparing feedback systems and their effectiveness in promoting learning in a particular context (i.e., NGSK learning in science). Moreover the research on formative assessment has largely been focused on the general student population and research on the achievement of sub-groups of students (e.g., gender, race, English Language Learners (ELL), students with an Individualized Education Program (IEP)) using these assessments is scarce (see Abedi, 2010; Elliot, Kettler, Bellow, Kurz, 2010). Thus to shed light on how these two types of formative feedback systems interact with learner characteristics, the analysis employed in this study specifically looked at achievement gains by sub-groups.

Method

Research Design

A quasi-experimental design comprising two groups was adopted in this study. The group that used the 1R Activity was referred to as the 1R Group and those that used the 4R Activity were referred to as the 4R Group in this analysis. Both the groups participated in the 10-week long curriculum on *Climate Change and Impacts on Ecosystems* (CCIE) (Songer et al., 2014) (see Chapter 1, this dissertation). The two groups engaged in the same embedded assessments at three time points in the curriculum (pre, mid, post). The difference between the two groups arose at the time of the embedded assessments, as both the groups did the CCIE curriculum (Songer et al., 2014) and identical pre and posttests.

Context of Study

This section provides a description of the assessment items and the details of the 1R and the 4R Activity respectively.

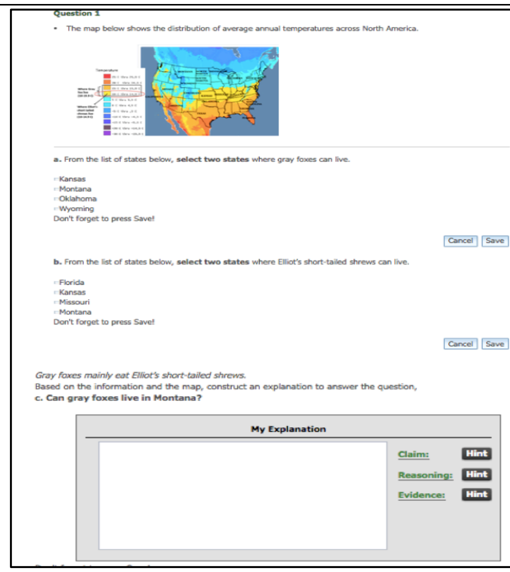
Assessment Items

The pre/post assessment items used in this study were curriculum sensitive and designed to measure *NextGen Science* Knowledge (NGSK). All assessment items had been previously validated and had been used to measure achievement gains in an earlier implementation cycle of the curriculum (Songer, Zaidi, Newstadt, in preparation). In an effort to avoid repetition on the pre, mid and posttest, surface level changes (e.g., changed name of a species) were made to two items on the midtests. Given the changes were only cosmetic, there was no threat to the construct validity of the items on the midtest (Messick, 1989). All tests (pre, mid and post) were embedded into the online CCIE curriculum (Songer et. al, 2014) and were administered online (see Appendix A for more information on the items and assessed NGSK). An example of an item and the assessed NGSK is shown in Table 2.1.

Table 2.1

A representative pre/post test item providing evidence of students' NGSK

<table>
<tr><td colspan="2">Assessed NGSK
Construction of a NGSK product by using a representation (temperature map) and information on a species and its prey's temperature ranges to show that the distribution of organisms is dependent on their environmental interactions both with other living things (biotic) and with nonliving things (abiotic).</td></tr>
<tr><td>Item on the pre/postest</td><td>Image of the pre/posttest item</td></tr>
<tr><td>Question 1
The map below shows the distribution of average annual temperature across North America.
a. From the list of states below, select two states where gray foxes can live.
• Kansas
• Montana
• Oklahoma
• Wyoming
Don't forget to press Save!

b. From the list of states below, select two states where Elliot's short-tailed shrews can live.
• Florida
• Kansas
• Missouri
• Montana
Don't forget to press Save!
Gray foxes mainly eat Elliot's short-tailed shrews.
Based on the information and the map, construct an explanation to answer the question,
c. Can gray foxes live in Montana?</td><td>
</td></tr>
</table>

As shown in Table 2.1, each assessment item had three parts presented as an item bundle (e.g., multiple items using similar stem information). The first two parts were designed to provide evidence of particular content knowledge (disciplinary core ideas) or skill (e.g., reading a map). These two parts were not scored and only the NGSK product (part c) was scored and used in the study.

The 1R Activity and the 4R Activity

As previously discussed the embedded formative assessments were administered three times during the curriculum as the *1R Activity* and the *4R Activity* (see Chapter 1, this dissertation) to the two groups of students respectively. These assessments occurred after a focal idea in the curriculum had been covered. The formative nature of these assessments made them assessments for learning where feedback was provided to students that informed their learning respectively. In an attempt to prevent students from viewing these assessments as summative assessments, the formative assessments were introduced as instructional activities (1R and 4R Activity) and words like "assessment" and "test" were not used to describe these formative assessments (see Ayala et al., 2008; Lucks 2003).

In the *1R Activity*, students responded online to a scientific question by constructing a NGSK product. The information gathered from students responses were then used to generate the most frequent score levels achieved by students on the midtest. The teacher and the author of this paper collaboratively came up with strategies to address the performance gap (gap between actual score level and maximum score level of 5). The teacher then discussed the strategies with the class in a classroom discussion format. This classroom level discussion was not customized to individual students' performance gaps but was focused on strategies pertaining to close the most common performance gaps observed in the classroom. The 1R Activity was designed in collaboration with the three participating teachers and two science coordinators of the school in which the 1R Activity was implemented. The 1R Activity reflected the existing formative assessment practices of the teachers in the 1R Group in which performance gaps were discussed with the students in the form of a classroom discussion. Thus in this aspect, the 1R Activity can be considered as a conventional formative assessment design, given it was similar to how the teachers administered formative assessments in their classrooms. The comparison between the 1R Activity and the 4R Activity is summarized in Table 2.2.

Table 2.2

Comparison of the formative assessment process in the 4R and the 1R groups[1]

Time	4R Group	1R Group
Day 1	Respond to the posed scientific question and submit the response.	Respond to the posed scientific question and submit the response.
	Provide and receive double-blinded peer review.	No peer review provided or received.
End of Day 1	Each individual receives additional review by an expert along with suggested guidelines for revision.	Expert review covering the most commonly occurring errors and gaps in learning provided to the teacher along with suggested instructional strategies to address the gaps in learning.
Day 2	A brief (10 minutes) teacher led class discussion informing the class of the observed gaps in learning.	A 40 minutes teacher led class discussion on gaps in learning along with suggested strategies to reduce the gap.
	Using customized feedback to revise and resubmit response.	No revision or resubmission required.
	Reflect on peer reviewing skills by reading the expert's review.	No reflection on reviewing.

Both the 1R and 4R Activities were grounded in the formative assessment framework proposed by Wiliam and Thompson (2007). In this framework Wiliam and Thompson conceptualize the following five key strategies for successful formative assessment:

(1) Clarifying and sharing learning intentions and criteria for success.

(2) Engineering effective classroom discussions and other learning tasks that elicit

[1] While the time allotted for the 4R/1R Activity was the same, the actual time spent on task by each individual student in the two groups could have differed.

evidence of student understanding.

(3) Providing feedback that moves learners forward.

(4) Activating students as instructional resources for one another, and

(5) Activating students as the owners of their own learning.

The 4R Activity (see Chapter 1, this dissertation) used the first two strategies in the *Respond* stage where students undertook the role of an author and constructed an NGSK product in response to a scientific question. The students were provided with the criteria for success in the form of a template outlining the different parts (i.e., claim, evidence and reasoning) that needed to be included in every NGSK product. The *Review* stage used the Wiliam and Thompson framework's third and fourth strategy of providing feedback to help learners move forward, and using students as instructional resources. Similar to the review process of many academic journals, *Review* adopted a double-blind process, where the identity of the author and the reviewer were hidden from each other. In the review process, the reviewer was provided with criteria to use for his/her evidence-based critique. Peer review was then made available to each author. As noted, an "expert" (scientist) also provided an additional review that complemented the peer review. The scientist then took on the role of an editor and used the reviews to provide the authors with guidelines for revision and resubmission. Students used the reviews they received to revise their NGSK products thereby advancing their own learning in the *Revise* stage. Lastly, in the *Reflect* stage, the students read the scientist's comments on their peer reviewing skills, and the scientist's review on the NGSK product that they had reviewed. They then engaged in a reflection activity on the peer review they have provided with an aim of improving the skill of peer reviewing in the future. This allowed them to be the owners of their own learning, as they engaged in a self-assessment of their own reviewing skills.

The 1R Activity also utilized the five strategies. Given that the *respond* phase in 1R is identical to the 4R Activity, the first two strategies were realized. The classroom discussion that focused on the observed learning gaps and provided strategies to address the gap made use of the third and fourth strategies. The participation of students in the discussion allowed them to become owners of their own learning, thereby employing the fifth strategy.

Study Participants and Settings

The students in this study were from the classrooms of eight teachers located across five schools in the U.S.A (see Table 2.3). All schools were publically funded and had implemented the curriculum in prior years. Assignment to the 1R and the 4R Group was done in the following way. From the pool of teachers that had implemented the curriculum in prior years, six schools were contacted to participate in the curriculum with the embedded 4R Activity. This resulted in four schools with five teachers teaching grades 6, 7, and 8 agreeing to implement the curriculum with the 4R Activity. Once the teachers volunteered to participate in implementing the curriculum with the 4R Activity, another three schools were contacted for assignment into the 1R Group. Out of the three schools contacted, two schools agreed to participate in the 1R Activity. However, only one of the two schools was chosen to implement the 1R Activity. School A was chosen for implementation of the 1R Activity because compared to the other school, which had only grade 6 and 7 students, School A had students from all grades (6-8th) participating in the curriculum. Additionally the number of students in School A was roughly equal to the total number students in schools implementing the 4R Activity. Choosing this school allowed for the two groups to be roughly balanced in terms of the number of participating students.

In total there were 257 students in the 1R Group and 230 students in the 4R Group that participated in the study. Students who had parental/guardian consent to be part of the study and had complete data (i.e., scores for pre, mid and posttests) were included in this study. These criteria resulted in the final sample consisting of 399 students from 6th, 7th and 8th grades. Out of the 399 students, 226 students were in the 1R Group and 173 students were in the 4R Group. Both the groups had a mix of 6th, 7th and 8th grade students. The distributions of students across the schools along with some school characteristics are shown in Table 2.3.

Table 2.3.

Participant and Setting Information

School ID	Formative Assessment Employed	U.S. Census Bureau Designated Divisions	% Enrollment in Free/Reduced Lunch Program	Grade	Participating Teachers	Participating Students
A	1R Activity	East South Central	50	6,7,8	3	226
B	4R Activity	New England	35	6	2	34
C	4R Activity	West North Central	30	7	1	33
D	4R Activity	East North Central	75	7	1	30
E	4R Activity	East North Central	12	6,7,8	1	76
				Total	8	399

Data Sources

The data used to address the research question were: (1) self-report of students' background information (e.g., race, gender); (2) teacher provided data on students classified as an English Language Learners (ELLs) and students with Individualized Educational Programs (IEPs); and (3) the scores on the NGSK products at pretest, mid-test and posttest time points. There were three assessment items in total on the pre and post-test, while each mid test had one item, which was similar (e.g., with only surface level changes made) or identical to the pre/posttest item. Given that there were three mid-tests administered, the total number of items for the mid time point was three.

Scoring of NGSK products

The scoring of NGSK products was done using a holistic rubric developed by Songer and colleagues (Songer, Zaidi & Newstadt, in preparation) that had been previously validated. The holistic rubric is reproduced with permission in Figure 2.1. Furthermore Appendix B provides information on the distinguishing characteristics between the score levels of the holistic rubric.

Level 5
Demonstration of a complete knowledge product that effectively incorporates relevant scientific terminology. *Elaboration:* Appropriate DCI and CC are used in the explanation/prediction, which are elaborated upon. The explanation/prediction includes a claim supported by relevant evidence along with appropriate reasoning that links the claim and evidence. AND NGSK product effectively incorporates relevant scientific terminology.
Level 4
Demonstration of a complete knowledge product but lacking effective incorporation of relevant scientific terminology. *Elaboration:* Appropriate DCI and CC are used in the explanation/prediction, which are elaborated upon. The explanation/prediction includes a claim supported by relevant evidence along with appropriate reasoning that links the claim and evidence. AND NGSK product does not incorporate any relevant scientific terminology, or uses only some of the relevant scientific terminology.
Level 3
Demonstration of a partially complete knowledge product that is appropriately elaborated upon and effectively incorporates some relevant scientific terminology. *Elaboration:* Appropriate DCI and/or CC are used in the explanation/prediction, which are elaborated upon. The explanation/prediction is partially complete as one of the three parts of the explanation (claim, reasoning, evidence) is missing or not explicitly stated. AND NGSK product effectively incorporates some relevant scientific terminology.
Level 2
Demonstration of a partially complete knowledge product that lacks elaboration and does not incorporate relevant scientific terminology. *Elaboration:* Appropriate DCI and/or CC are used in the explanation/prediction, which are not adequately elaborated upon. The explanation/prediction is partially complete as two of the three parts of the explanation (claim, reasoning, evidence) are missing or not explicitly stated. AND NGSK product does not effectively incorporate relevant scientific terminology.
Level 1
Demonstration of an incomplete knowledge product that lacks elaboration and does not incorporate relevant scientific terminology. *Elaboration:* Only uses appropriate DCI, which is not adequately elaborated upon. AND NGSK product does not effectively incorporate relevant scientific terminology.
Level 0
Demonstration of an inaccurate knowledge product. *Elaboration:* Inappropriate DCI or scientifically inaccurate use of the DCI

Figure 2.1. Holistic rubric used for scoring explanations/predictions. Reproduced with permission from "Classroom-based holistic assessment and evidence of middle school students' Next Generation Science Knowledge", by N.B. Songer, S.Z. Zaidi & M.R. Newstadt, in preparation.

The consensus estimates method (Stemler, 2004) was adopted to establish reliability. In accordance with this method, another rater was trained in interpreting the holistic rubric followed by an independent scoring session of ten identical NGSK products by the first author of this paper and the rater. The process of scoring and comparing scores was repeated until an inter-rater agreement of 90% was established for each of the NGSK products. Additionally the Cohen's Kappa (Cohen, 1960) statistic was also used to establish inter-rater reliability. The Kappa value was 0.82, which is considered as a high degree of agreement among raters (Cohen, 1960). Following this, the first author of this paper scored all NGSK products with spot checks performed by the other rater on 10% of the data.

The scores at the mid-test time point for the 4R Group were the scores of the original NGSK product and not the revised product. This was because the students in the 1R Group did not revise and resubmit the NGSK product. Using the scores on the original NGSK product resulted in the "treatment effect" to be between the mid and post time points and not between the pre and midtest time point. Additionally, all three midtests were considered as representing one intermediate time point between pretest and posttest. This was because there was a "treatment effect" after the administration of the first midtest. By considering the time between the first midtest and posttest as the difference in treatment allowed for comparing the achievement outcomes pertaining to students in the 1R and the 4R groups. By considering the three midtests as one intermediate time point also allowed for the pre, mid and post tests to have the same score range of 0-15 levels (referred to as points in this study), with each question having a score range of 0 to 5 points.

Check for Equality of Distribution

In order to establish that there was no bias in sampling students for the two groups, the Kolmogorov-Smirnov test (Massey, 1951) was conducted to check for the equality of distribution of the pretest scores. The test revealed that the null hypothesis (i.e., there is no statistical difference in the distribution of the pretest scores for the two groups) could not be rejected even at a significance level of 0.10 (the p-value for the test was 0.236). This result showed that the two groups were sampled from the same larger population with a similar distribution for the pretest scores.

Analysis

In order to compare the achievement scores of students on NGSK products pre, mid and post time points, three approaches to analysis were employed. First, this study examined the extent of the overall achievement gains for the two groups, the 4R Group and the 1R Group respectively. Here, overall achievement gains referred to looking at the score gains from pre-to-mid time point and from mid-to-post time point where the scores of the three separate mid-tests were combined and treated as one single test administered between the pre and post-test. It also explored the achievement gains for different sub-groups of population, such as female, racial-ethnic minorities, English language learners (ELL) and students with Individualized Education Programs (IEP) were analyzed to determine the variability in the effects of group assignment on these attributes. To understand the differences in achievement gains based on the difficulty of the item, a second step in analysis was to conduct an item level analysis for the three NGSK items. This item level analysis examined the performance gap between the 1R and the 4R Group for three items of varying difficulties.

As a final analysis step, the percentage distribution of students was examined at different pretest scores and posttest scores for the two groups. This analysis intended to showcase the percentage distribution of students that began at a specific pretest score range and ended at a particular post-performance level.

Before conducting the analysis described above, dimensionality analysis of the three NGSK items as well the difficulty of the three NGSK items relative to each other was determined using procedures of Item Response Theory (Hambleton & Swaminathan, 1985; Hambleton, Swaminathan, & Rogers, 1991). Dimensionality analysis provided information on the number of underlying dimensions that were needed to explain the covariance between the items. Information on the difficulty analysis allowed for the interpretation of the results in light of the difficulty of the items.

Estimating the Difficulty Coefficients of the three NGSK Items

The data used to determine the difficulty of the items was from a previous implementation cycle (N=258) that did not use any of the embedded assessments (1R or 4R) described in this paper. In the previous implementation cycle, 90% of the data came from school sites that were used in this study. The difficulty of the three items was estimated on the posttest data (i.e., after the students had completed the curriculum).

First a dimensionality analysis was conducted to determine the number of dimensions the three NGSK items were measuring. Dimensionality checks were done using exploratory factor analysis. In this analysis factors with eigen values above 1 were retained. Dimensionality analysis revealed that there was only one factor with eigen value above 1 thereby establishing that the three questions were best fit by a single factor and therefore represented a unidimensional construct. The difficulty of the items was determined using Item Response Theory (Hambleton & Swaminathan, 1985; Hambleton, Swaminathan, & Rogers, 1991) where a one parameter logistic Partial Credit Model (Masters & Wright, 1997) was estimated. This was done through the Stata (Stata Corp, 2011) program *gllamm* (Rabe-Hesketh, Skrondal & Pickles, 2004), which estimated the parameters using marginal likelihood and performed adaptive quadrature. The difficulty coefficient revealed that the most difficult item was item 2 (average item difficulty =0.83), followed by item 3 (average item difficulty =0.57), and the least difficult of the three items was item 1 (average item difficulty =0.28). In other words, item 2 was three times as difficult as item 1, while item 3 was twice as difficult as item 1.

Fixed Effects Model to Examine Overall Learning

To address the extent of the difference in achievement for the two groups, a multiple linear regression model with fixed effects was estimated in *Stata* (StataCorp, 2011). The fixed effects model allowed for unobserved individual heterogeneity, which meant that it controlled for all time-invariant variables (e.g., gender, race, geographic location) that did not change in the period being studied. The equation representing the fixed effects regression model for individuals *i* at time *t* is as follows:

$$Y_{it} = \alpha_i + \mu_t + \beta X_{it} + \varepsilon_{it} \qquad \text{(eq 1)}$$

Where:

- Y_{it} is the outcome variable for individuals at each time period
- α_i is the intercept for student fixed effects
- μ_t is the intercept for time fixed effects
- X_{it} are the time variant independent variables
- β is the coefficient of the independent variable
- ε_{it} is the error term that is statistically independent of all other variables

The fixed effects model was specified at the student level. In the model the outcome variable was the total score for each individual at the three time points (pre-, mid- and post) respectively. The total score at each time point had a range of 0 to 15 points. Predictor variables used in the fixed effects model were the time-varying variable that indicated the time point at which the test was administered. For this, dummy variables were created to indicate pretest and posttest time points. Thus the comparison across time was made with respect to the intermediate time point.

Given that the fixed effects model controls for all time-invariant predictors (such as group assignment, gender, race, etc.) and does not estimate the coefficients for these time-invariant predictors, the model was expanded to include the interaction effects of these time-invariant predictors with the time point at which the test was given. This analysis was done to understand the performance of students in the two groups and those students that have been previously documented to benefit from formative assessment interventions. For this model, interactions between the time varying variables (time point of test administration) and time-invariant variables were included. The time-invariant variables that were considered for interaction effects were, (1) group assignment, (2) grade of the student (3) race of the student (categories include White, African American, and Other races), (4) gender, (5) whether the student was an English Language Learner (ELL), and (5) whether the student was in an Individualized Education Program (IEP).

A fixed effects model with two-way interactions between group assignment and time point of test administration was created. This allowed for determining of the effect of group assignment on score gains for all students. The equation of the model for student *i* at time point *t* was as follows:

$$\begin{aligned}\text{TotalScore}_{it} = {} & \alpha_i + \mu_t + \beta_1\, \text{Pre}_{imid} + \beta_2\, \text{Post}_{imid} + \beta_{1i}\, (\text{Pre x Group}) + \\ & \beta_{2i}\, (\text{Post x Group}) + \varepsilon_{it}\end{aligned} \qquad \text{(eq2)}$$

The standard errors of the fixed effects model were clustered at the classroom level, which implied that standard errors for observations for a teacher's classroom could be correlated.

With an aim to explore the existence of heterogeneity in the effect of the 4R Activity the main model (equation 2) was expanded to include three-way interaction effects. In order to conserve the degrees of freedom in the model, interaction effects with large p-values ($p > 0.80$) and very small coefficients ($<|0.1|$) were removed. The interaction terms that were eliminated from the model were the interactions effects with IEP and grade. The model with three-way interaction effects had interaction effects of group, gender, race, and ELL status. Three-way interaction effects allowed in exploring the existence of heterogeneity in the effect of the 4R Activity based on gender, race and ELL status for pre-to-mid and mid-to-post times. The equation of the model for student *i* at time point *t* was as follows:

$$\begin{aligned}\text{TotalScore}_{it} = {} & \alpha_i + \mu_t + \beta_1\,\text{Pre}_{imid} + \beta_2\,\text{Post}_{imid} + \beta_{1i}\,(\text{Group x Pre}) + \beta_{1i}\,(\text{Gender x Pre}) + \\ & \beta_{1i}\,(\text{Race x Pre}) + \beta_{1i}\,(\text{ELL x Pre}) + \beta_{1i}\,(\text{Gender x Pre x Group}) + \\ & \beta_{1i}\,(\text{Race x Pre x Group}) + \beta_{1i}\,(\text{ELL x Pre x Group}) + \beta_{2i}\,(\text{Group x Post}) + \\ & \beta_{2i}\,(\text{Gender x Post}) + \beta_{2i}\,(\text{Race x Post}) + \beta_{2i}\,(\text{ELL x Post}) + \\ & \beta_{2i}\,(\text{Gender x Post x Group}) + \beta_{2i}\,(\text{Race x Post x Group}) + \\ & \beta_{2i}\,(\text{ELL x Post x Group}) + \varepsilon_{it} \qquad \text{(eq3)}\end{aligned}$$

Results

Two-way Interaction Effects

The results for the fixed effects model with two-way interaction effects of group assignment are shown in Table 2.4.

Table 2.4

Fixed effects model with two-way interaction terms for total score over time

Variables	Coefficient	Robust Standard Error	P > |t|
Test administration time point (Reference group is mid time)			
Pre test time point	**-3.50**	**0.14**	**<0.0001*****
Post test time point	**1.96**	**0.05**	**<0.0001*****
Two-way Interaction Terms			
Group (reference is 1R Group)			
4R Group x Pre	0.07	0.20	0.71
4R Group x Post	**2.29**	**0.43**	**0.001****

Note: ***significant at p<0.001; ** significant at p<0.05

As seen in Table 2.4 there were significant gains for all students regardless of group assignment from pre-to-mid and from mid-to-post time points. The score gain from pre-to-mid was 3.50 points (SD =2.12) while the mid-to-post gain was 1.96 points (SD=1.96). Given the study's primary interest was to examine the degree to which the score gains differed by group assignment, the interpretation of the two-way interaction effects is more relevant here.

The interaction of group assignment with the time point of test administration revealed that the difference in achievement gains of the two groups (1R and 4R) was significant at the mid-to-post time (p<0.05). Students in the 4R Group attained 2.29 points (SD=1.99) higher than 1R Group students. In other words students in the 4R

Group had a score gain of 4.25 points (SD=1.99) from mid-to-post while the 1R Group students had a score gains of only 1.96 points (SD=1.21) in the same time period.

Three-way Interaction Effects

The interaction of group assignment with time point of test administration indicated that all students in the 4R Group had a 2.29 points (SD=1.99) higher score gain than 1R Group students. To explore the achievement gains of specific sub-populations (e.g., females, African Americans) in the 4R Group fixed effects models with three-way interactions was used. The results for the fixed effects model with the three-way interaction effects are shown in Table 2.5. The coefficients that were significant (at $p<0.001$; $p<0.05$; $p<0.10$) have been highlighted using a bold font.

Table 2.5
Fixed effects model with three-way interaction terms for total score over time

Variables	Coefficient	Robust Standard Error	P > |t|
Test administration time point (Reference group is mid time)			
Pretest time point	**-3.66**	**0.09**	**<0.0001*****
Posttest time point	**1.96**	**0031**	**<0.0001*****
Two-way Interaction Terms			
Group (reference group is 1R Group)			
4R Group x Pre	-0.02	0.29	0.94
4R Group x Post	**1.40**	**0.13**	**<0.0001*****
Gender (reference group is Male)			
Female x Pre	0.51	0.33	0.16
Female x Post	0.05	0.15	0.77
Race (reference group is White)			
African American x Pre	-0.27	1.10	0.82
African American x Post	0.33	0.20	0.14
Other Races x Pre	-0.29	0.37	0.46
Other Races x Post	-0.22	0.16	0.21
English Language Learners (reference group is non ELLs)			
ELL x Pre	1.21	0.83	0.19
ELL x Post	0.59	0.39	0.19
Three-way Interaction Terms			
Gender (Reference groups are Male, Mid, 1R Group)			
Female x Pre x 4R Group	-0.35	0.48	0.49
Female x Post x 4R Group	**0.80**	**0.21**	**0.007****
Race (Reference groups are White, Mid, 1R Group)			
African American x Pre x 4R Group	0.90	1.14	0.45
African American x Post x 4R Group	**0.82**	**0.29**	**0.03****
Other Races x Pre x 4R Group	0.67	0.74	0.40
Other Races x Post x 4R Group	**0.77**	**0.43**	**0.08***
English Language Learners (Reference groups are non ELLs, Mid, 1R Group)			
ELL x Pre x 4R Group	-1.47	1.19	0.26
ELL x Post x 4R Group	0.37	0.72	0.62

Note: ***significant at p<0.001; ** significant at p<0.05; * significant at p<0.10

The statistically significant two-way and three-way interactions effects in Table 2.5 are most meaningful to interpret given they examine the degree to which the achievement gains of students in different sub-groups differed with respect to group assignment. The two-way interaction effect of group assignment with test time period indicated that non-ELL, white males in the 4R Group had a 1.40 points higher gain in scores compared to non-ELL, white males in the 1R Group. In other words, the score gains of these students were 3.36 points (SD=2.00) while their counterparts in the 1R Group had a score gain of 1.96 points (SD=1.14) in the mid-to-post period.

The three-way interactions explored the heterogeneity of the 4R Activity's effect on score gains of the sub-groups. The statistically significant interaction terms indicated that there existed heterogeneity in the effect of the 4R Activity on being female, African American and being from the "Other Races" category for the mid-to-post period. The positive coefficients indicated that in addition to the effect of 4R Group on score gains (1.40 points), there was an additional score gain equal to the coefficient of that three-way interaction for the mid-to-post period. For example, non-ELL white females in the 4R Group had a 0.80 points increase in the mid-to-post time point. This made their score increase in mid-to-post period to 4.16 points (1.96+1.40+0.80) (SD=1.51). Similarly, being African American in the 4R Group resulted in an additional 0.82 points increase in score in the mid-to-post period, while being from the "other races" category had a 0.77 points increase in addition to the 1.40 points increase as a result of being in the 4R Group. Thus, non-ELL African American males had a score increase of 4.18 points (SD=1.39), while non-ELL

African American females had a score increase of 4.98 points (SD=1.45) in the mid-to-post period. Furthermore, non-ELL other race male students had a score gain of 4.13 points (SD=1.95) while non-ELL other race female students had a score gain of 4.93 points (SD=2.55) in the mid-to-post period in the 4R Group.

The three-way interaction effects for students classified as ELLs were not significant. This showed that there was no heterogeneity in the 4R Activity's effect for ELL status. In order to explore the difference in the achievement gains of students with ELLs, Fisher's exact test were conducted on ELLs in the 1R Group and the 4R Group. Fisher's exact test was determined the difference in the achievement gains of ELLs in pre-to-mid and mid-to-post time periods for the 4R and 1R groups respectively. Fisher's exact test for pre-to-mid period was statistically insignificant ($p=0.681$). For the mid-to-post period, Fisher's exact test was statistically significant ($p=0.09$) indicating that there was a statistically significant relationship between group assignment and achievement gains in mid-to-post time periods. The magnitude of achievement gains in the two time periods is shown in Table 2.6. The table showcases the increase in score from the mid-to-post time point for ELL students in the 1R and the 4R group respectively.

Table 2.6

Achievement gains from mid to post time point for students classified as an English Language Learner (ELL)

Midscore Points	ELL Students in 1R Group (N=16)	ELL Students in 4R Group (N=21)
0-5 points	4	6
6-10 points	10	13
11-15 points	2	2
Gain in Points from mid-to-post Point		
Points Increase in mid-to-post time	ELL Students in 1R Group	ELL Students in 4R Group
0-2 points	10	2
3-5 points	5	9
6-8 points	1	10

Note. ELL = English Language Learner

From Table 2.6 it is seen that in the 1R group the majority of students classified as ELLs demonstrated a score increase of 0-2 points. In contrast, the majority of students classified as ELLs in the 4R Group showed a score increase of 6-8 points.

Achievement Gains of Students with IEPs

As discussed earlier, the interaction term for students with IEPs was not included in the model because of very small (< |0.01|) and statistically insignificant coefficient (p=0.6) of the interaction effects of group assignment with IEP status. The sample size of IEP students was also small (n=16 in the 4R Group; n=19 in the 1R Group). Similar to conducting Fisher's exact test for ELLs, Fisher's exact test was conducted for students with IEPs for the pre-to-mid and mid-to-post periods. Fisher's exact test for the pre-to-mid period was statistically insignificant (p=0.66) while it was statistically significant for the mid-to-post period (p=0.03). The results of Fisher's exact test imply that there was a difference in the achievement gains of students in the two groups in the mid-to-post period. For the purpose of comparison of achievement gains of IEP students in the two groups, achievement gains from mid-to-post of IEP students are presented in Table 2.7. This table provides a quick comparison of the achievement gains of IEP students that fell in the different score increase categories.

Table 2.7

Achievement gains from mid to post time point for students with an Individualized Education Program (IEP)

Midscore Points	IEP Students in 1R Group (N=19)	IEP Students in 4R Group (N=16)
0-5 points	9	10
6-10 points	10	6
11-15 points	0	0
Gain in Points from mid-to-post Point		
Points Increase in mid-to-post time	IEP Students in 1R Group	IEP Students in 4R Group
0-2 points	12	1
3-5 points	7	11
6-8 points	0	4

Note. IEP = Individualized Education Program

Table 2.7 shows that students with IEPs in the 1R Group and the 4R Group had comparable scores at the mid test. However on the posttest, majority of students with IEPs in the 1R Group demonstrated a 0-2 points increase in score from the midtest to the posttest. On the other hand, students with an IEP in the 4R Group showed an increase in 3-5 points from midtest to posttest. Additionally, there were students that showed a 6-8 points increase in their score from the midtest to the posttest.

Fixed effects model for each item on the assessment

The analysis presented above provides information on achievement gains for the overall test. This section looks at the analysis that breaks the achievement gains down by each NGSK item. The difference in achievement gains of students within the two groups was investigated using the difficulty of any given item relative to the others. For this analysis only time-varying variables were considered and no interaction terms were estimated since the purpose of this analysis was to compare achievement gains by item for the two groups. Table 2.8 shows the estimates for item 1 along with the interaction of the time points with the group assignment (4R and 1R Group).

Table 2.8

Fixed effects model for score of item 1 over time

Variables	Coefficient	Robust Standard Error	P > |t|
Test administration time point (Reference group is mid time)			
Pre time	-1.21	0.06	<0.0001
Post time	0.75	0.04	<0.0001
Interaction Terms			
Group (reference is 1R Group)			
4R Group x Pre	0.24	0.11	0.026
4R Group x Post	0.61	0.09	<0.0001

Note: All coefficients are significant at $p<0.05$

In Table 2.8 it is seen that for item 1, which was the easiest of the three items, there was a 1.21 points increase in the score from pre-to-mid test for all students and this increase is highly significant. Additionally, the increase in scores from mid-to-post for all students was 0.75 points and this increase is also highly significant ($p<0.0001$).

When looking at the increase in score based on group assignment, for the pre-to-mid point, students in the 4R Group gained 0.24 points more than the 1R Group students. However, this increase is less than one-fourth of a point and its practical significance is low. For the mid-to-post time point the score increase by the 4R Group students was 0.61 points greater than that of the 1R Group students. This increase is slightly greater than the increase observed in the pre-to-mid time frame. The achievement gain of students in the mid-to-post time frame for the 4R Group was 1.36 points (SD=1.06), compared to 0.75

points (SD=0.61) for the 1R Group students. Thus 4R Group's achievement gains were 1.8 times that of 1R Group's achievement gains.

Proceeding with examining the achievement gains for item 2, Table 2.9 shows the results of the fixed effects model estimated for this question.

Table 2.9

Fixed effects model for score of item 2 over time

Variables	Coefficient	Robust Standard Error	P > |t|
Test administration time point (Reference group is mid time)			
Pre time	-0.82	0.11	<0.0001
Post time	0.57	0.03	<0.0001
Interaction Terms			
Group (reference is 1R Group)			
4R Group x Pre	-0.33	0.12	0.03
4R Group x Post	0.93	0.08	<0.0001

Note: All coefficients are significant at $p<0.05$

The NGSK item 2 was the most difficult of the three NGSK items (based on item difficulty indices). As shown in Table 2.8, for item 2 a 0.82 points increase in score occurs for all students in the pre-to-mid time, while for mid-to-post time a 0.57 points increase was observed for the students. Both these coefficients are highly significant ($p<0.0001$). When broken down by group assignment, the score difference in pre-to-mid score between the 4R and 1R group was 0.33 points. Thus the 4R Group students showed an increase of 1.15 points compared to an increase of 0.82 points by the 1R Group during pre-to-mid time.

At the mid-to-post time point the 4R Group students had scored 0.93 points higher than the 1R Group students. The score gain for the 4R Group students was 1.5 points

(0.57+0.93) (SD=1.99), while for the 1R Group students it was 0.57 points (SD=1.21). The 4R Group students' gain in scores was 2.6 times that of the 1R Group students.

For the last item, the fixed effects regression results are shown in Table 2.10. For the third item, which was twice as hard as the first question (based on item difficulty indices), all students showed an increase in score by 1.46 points from pre-to-mid time.

Table 2.10

Fixed effects model for score of item 3 over time

Variables	Coefficient	Robust Standard Error	P > |t|
Test administration time point (Reference group is mid time)			
Pre time	**-1.46**	**0.06**	**<0.0001**
Post time	**0.65**	**0.07**	**<0.0001**
Interaction Terms			
Group (reference is 1R Group)			
4R Group x Pre	0.17	0.84	0.43
4R Group x Post	**0.75**	**0.30**	**0.04**

Note: Coefficients in bold font are significant at $p<0.05$

The interaction of group with pre time point was not significant at $p<0.05$ level thereby suggesting that there was no difference in the achievement gains during pre-to-mid for the two groups. For mid-to-post time, the increase in scores for the 4R Group was 1.4 (0.75+0.65) points (SD=1.08), while it was 0.65 points (SD=0.65) for the 1R Group students. Thus the 4R Group students showed an achievement gain that was 2.2 times that of the 1R Group students.

Achievement gains based on pretest scores

As a final step in the analysis, the achievement gains of students based on their pretest scores was performed. First the range of scores achieved by students at the pretest and posttest was determined. On the pretest, student scores ranged from 0 to10 points on a 15 points test. These upper and lower bounds of the achieved scores were used to define three pre-performance levels. The first group consisted of a score range of 0-3 points, the second was from 4-7 points and third range was from 8-10 points. These pretest score ranges functioned as the baseline levels. A chi-squared test for difference in proportions in the pretest score ranges between the 1R and 4R groups was performed at the 5% significance level. The result indicated that there was no statistical difference between the proportion of students in three pretest score ranges for the 1R and 4R group respectively (Pearson chi-squared (2, N=399)=4.89; p=0.087). Thus, the proportion of students in pretest score ranges was comparable for the 1R and the 4R group.

Similar to the pretest score ranges creation, post-performance levels were created using the minimum and maximum score range achieved at the posttest. At the posttest, the minimum score achieved was 3 points and the maximum score was 15 points. This score range was divided into three post-performance levels with a score range of 3-7 points, 8-12 points and 13-15 points respectively. The 3-7 points range captured scores that were below 50% of the total points on the test and was referred to as the "*below 50%"* post-performance level. Similarly, the 8-12 points range was referred to as the "*between 50-80%"* post-performance level while the 13-15 score range was referred to as the "*above 80%"* post-performance level. Similar to the chi-squared test on the pretest score ranges, a chi-squared test for difference in proportions in the post-performance levels between the 1R and 4R groups was also performed at the 5% significance level. The results indicated that there was a statistically significant difference in the proportions of students in the three post-performance levels for the 1R and 4R group respectively (Pearson chi-squared (2, N (399)=81.54; p<0.0001). Unlike the pretest score ranges, there was a difference in the proportion of students in the three post-performance levels for the 1R and the 4R group. The chi-squared tests for difference in proportion on the pretest score ranges and post-performance levels for the two groups indicated that there was no statistical difference in proportion for the pretest score ranges. However, the proportion of students in the three post-performance levels was different for the 1R and the 4R group.

Given the different proportion of students in the post-performance levels for the 1R and the 4R group, Table 2.11 was created to represent the percentage distribution of students in each group (1R, 4R) as well as the differences in percentages of students in the two groups at the post-performance levels.

Table 2.11

Difference in the percentage of students in post-performance levels in the 4R Group and the 1R Group

4R Group (N=173)					
Pretest score ranges (pts.)	Post-Performance Levels Based on Posttest Scores				
		Below 50% (3-7 pts.) (%)	Between 50%-80% (8-12 pts.) (%)	Above 80% (13-15 pts.) (%)	Row Total (%)
	0-3	9	35	11	55
	4-7	1	19	19	39
	8-10	0	1	5	6
Column Total (%)		10	55	35	100
1R Group (N=226)					
Pretest score ranges (pts.)	Post-Performance Levels Based on Posttest Scores				
		Below 50% (3-7 pts.) (%)	Between 50%-80% (8-12 pts.) (%)	Above 80% (13-15 pts.) (%)	Row Total (%)
	0-3	31	35	0.5	66.5
	4-7	4	24	3	31
	8-10	0	1.5	1	2.5
Column Total (%)		35	60.5	4.5	100
Difference between 4R & 1R					
Pretest score ranges (pts.)	Post-Performance Levels Based on Posttest Scores				
		Below 50% (3-7 pts.) (%)	Between 50%-80% (8-12 pts.) (%)	Above 80% (13-15 pts.) (%)	
	0-3	-22	0	+10.5	
	4-7	-3	-5	+16	
	8-10	0	-1	+4	

Note. + Indicates higher % students in 4R Group compared to 1R Group
- Indicates lower % students in 4R Group compared to 1R Group

In Table 2.11, the major point of departure with respect to the proportion of students in the post levels was for the *below 50%* and the *above 80%* performance levels (see “column total” in Table 2.11). A total of 35% of the students in the 4R Group were in the *above 80%* post-performance level, while only 4.5% of students were in this level in the 1R Group. For the *between 50-80%* post-performance level there were 55% of students from the 4R Group were in this level, while 60.5 % of 1R Group students were in this level. For the *below 50%* post-performance level there were only 10% of students in the 4R Group while 35% of the 1R Group students were in the *below 50%* post-performance level. Thus, the 4R Group saw an increase in the percentage of students in the *above 80%* level and a decrease in the percentage of students in the *below 50%* post-performance level.

Examining the difference in percentage distribution of students in the 1R Group and the 4R Group, it was observed that the major point of departure between the two groups were for the *below 50%* and the *above 80%* post-performance levels. In the 4R Group, only 9% of students fell in the *below 50%* post-performance level which was in contrast to the 1R Group where 31% of the students fell in the *below 50%* post-performance level. This resulted in 22% fewer students in the 4R Group who began at the 0-3 points pretest score range and achieved a *below 50%* post-performance level. In the *above 80%* post-performance level, there were 10.5% more students in the 4R Group that began at the 0-3 points pretest and achieved an above 80% post-performance level. Furthermore there were 16% more students in the 4R Group than the 1R Group that began at the pretest score range of 4-7 points and achieved an *above 80%* post-performance level.

Discussion

The findings from this study indicate that formative assessments that use customized feedback and provide opportunities to revise (i.e., 4R Activity) demonstrate higher achievement gains for all students on the measured NGSK. The fixed effects regression analysis indicated that group assignment (individualized [4R Activity] or collective [1R Activity]) was the strongest predictor of achievement gains. In other words, the type of formative assessment administered had the largest impact on student achievement. The findings pertaining to group assignment revealed that all students in the 4R Activity demonstrated score gains that were 2.29 points higher than the students in the collective feedback system (1R Activity). In other words when the 1R Group students demonstrated a gain of 1.96 points in the mid-to-post period, the 4R Group students demonstrated a gain of 4.25 points on a 15-points scale. Furthermore, in addition to the gains in score as a result of being in the 4R Group, there were additional score gains for non-ELL females and for non-ELL non-white racial/ethnic groups (e.g., African Americans, Other Races). This resulted in non-ELL female students from non-white racial/ethnic groups in the 4R Group to show the largest gains in score in the mid-to-post period (e.g., 4.98 points increase for African American females compared to 3.36 points increase for white males in the 4R Group). These findings suggest the potential of the 4R Activity in supporting the learning of students, with particular improvements for those underrepresented in science and engineering.

With respect to ELLs and students with IEPs, an exploratory analysis using Fisher's exact test revealed larger learning gains in the 4R Group for these student sub-groups. However, the small sample size of students with IEPs and ELLs considerably limits the power of this finding. Nevertheless, this result serves as an important step for future explorations of achievement of students in IEPs and ELL students in such iterative formative assessments.

These findings are significant for many reasons beyond their statistical significance. First, from a perspective of employing such interventions in classroom, sub-group achievement gains as well as overall achievement gains based on the form of formative assessment administered show that learning and therefore achievement can be accelerated with the use of formative assessments like the 4R Activity. Second, the sub-group findings showcase the gains in achievement for specific sub-groups. These findings are relevant because they look at how a particular assessment system functions to promote achievement of learners that have been historically underrepresented in postsecondary science, technology, engineering and mathematics (STEM) education and STEM careers (see Burke & Mattis, 2007; NRC 2011). This in turn allows educators to use such assessments fruitfully to enhance achievement of all students instead of employing interventions that might be more suitable for only the dominant (white male) student groups. These results help in showcasing the usability of the 4R Activity in classroom environments focused on equitable learning opportunities for all students.

When examining achievement based on the difficulty of the items on the assessment, the item-level analysis revealed that with increase in difficulty of the question, the performance gap between students in the two forms of formative assessment increased. In other words, the 4R Activity was most effective for difficult items, with the largest performance gap between the two groups occurring for the most difficult question (the 4R Group's achievement gain was 2.6 times that of the 1R Group). This analysis is important when considering the placement and frequency of formative assessments like the 4R Activity. Frequent administration of these assessments are desirable but might not be a realistic goal in actual classrooms, given they are more labor and time intensive.

However, these results serve as a basis on which the decision to administer these assessments can be made. The findings reveal that these formal assessment opportunities might be most fruitful for more difficult concepts, as the performance gap between the two feedback systems was the largest for the most difficulty item.

Another significant finding was the differential achievement gains based on gender and race in the 4R Activity with greater achievement gains observed for traditionally underrepresented populations in STEM (e.g., females, non-white racial/ethnic groups) than their white male students. One of the potential attributes to differential achievement gains could be the double-blinded review process in the 4R Activity. The anonymous aspect removes the threat of gender bias and other inter-person dynamics that might otherwise bias the review process.

It is also important to note some lagged short term effects that are seen as improved performance of the 4R Group students in pre-to-mid time points. One potential reason for this might be that the 4R Group students had already done the activity once when they took their second mid-test. This could have contributed towards their increased performance in the second midtest. However we see that at the third midtest question, the performance gap between the two groups is once again not statistically significant. This implies that the students in the 1R Group catch up with the 4R Group students by this time point. Even though these short-term effects are small and fade out quickly they pave the way for future research on ways to enhance and sustain short-term effects.

The final step in analysis pertained to documenting the difference in the two groups based on the percentage distribution of students in the post-performance levels. These pre-to-post shifts helped us examine the growth of students that began at different pretest score ranges. This analysis showed that many more students (35%) in the 4R Group were able to move to achieving a score of 80% or higher than students the 1R Group (4.5%). Additionally students who achieved a score of less than 50% on the posttest, considerably decreased in the 4R Group (10%). The analysis revealed that the 1R Group students have made smaller achievement gains as compared to the 4R Group. One way to explain this outcome is that perhaps the 4R Activity with customized feedback and explicit opportunity to revise helped in accelerating rate of learning and therefore students had higher achievement gains. As a future step, it will be beneficial to systematically study if these effects of accelerated learning are sustained over a long period of time and if students are able to successfully apply this acquired learning to similar focal ideas in science.

Given that the type of formative assessment administered was a strong predictor of achievement gains, it is essential to discuss some of the unique features of the 4R Activity that may have contributed to the gains. Building on the affordances of performance-based assessments where opportunities to revise are provided in an effort to build proficiency, the 4R Activity adopted this strategy within the formative assessment. The two unique features that were incorporated in the 4R Activity were: (1) anonymous peer and expert feedback that provided customized feedback to each student, and (2) opportunity to revise with guided instructions on how to undertake the revision process. The customized feedback from a peer and an expert complemented each other and catered to the needs of diverse learners in the classroom. The opportunity to revise allowed the students to explicitly act on the feedback to close the performance gap between their actual and desired level of performance.

This study claims that both these features contributed in positively impacting student performance. The claim is grounded in research that has shown the importance of revision in achieving proficiency in a domain and in closing the learning gap (e.g., Nichols &Berliner, 2007; Hughes, 2010). Research on the use of performance-based assessments to build proficiency in an area has shown that opportunities provided to revise supports student learning in that area (Gallagher, 2007; Wiggins, 2006; Nichols &Berliner, 2007). Additionally, Hughes (2010) work suggests that providing opportunities to use corrective feedback is especially beneficial for low- achieving students.

Another contributor to learning for the 4R Group could be the use of a system that simulated the peer review process of academic journal. Student motivation could be higher when using this system as classroom observations revealed that students expressed increased engagement because they felt that they were using a system similar to what professional scientists, engineers and academics used. Additionally, there might have been extrinsic motivation to do the activity as classroom observations also revealed that students, in general were excited that a scientist at a university read their work and commented on it. In line with Yin and colleagues (2008) hypothesis of looking at formative assessments in relation to enhanced motivation, future research needs to examine the effectiveness of formative assessment interventions in light of motivation as a factor.

References

Abedi, J. (2010). Research and recommendations for formative assessment with english language learners. In Andrade, H., & Cizek, G. J. (Eds.). (2010). *Handbook of formative assessment*. Routledge.

Andrade, H.L. (2010). Students as the definitive source of formative assessment: Academic self-assessment and the self-regulation of learning. In Andrade, H., & Cizek, G. J. (Eds.). (2010). *Handbook of formative assessment*. Routledge.

Ayala, C. C., Shavelson, R. J., Araceli Ruiz-Primo, M., Brandon, P. R., Yin, Y., Furtak, E. M., & Tomita, M. K. (2008). From formal embedded assessments to reflective lessons: The development of formative assessment studies. *Applied Measurement in Education*, *21*(4), 315-334.

Azevedo, R., & Bernard, R. M. (1995). A meta-analysis of the effects of feedback in computer-based instruction. *Journal of Educational Computing Research, 13*(2), 111–127.

Bell, B., & Cowie, B. (2001). *Formative assessment and science education.* London:Kluwer.

Burke, R.J., and Mattis, M.C. (Eds.). (2007). *Woman and minorities in science technology, engineering, and mathematics: Upping the numbers.* Northampton, MA: Edward Elgar.

Black, P., & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education:Principles, Policy and Practice, 5*(1), 7–73.

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

Bloom, B.S., Hastings, J.T., & Madaus, G.F. (Eds). (1971). *Handbook of formative and summative evaluation of student learning*. New York: McGraw-Hill.

Brown, S., Rust, C., & Gibbs, G. (1994). *Strategies for diversifying assessment in higher education*. Oxford: Oxford Centre for Staff Development.

Cohen J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37-46.

Elliott, S.N., Kettler, R.J., Beddow, P.A., & Kurz, A. (2010). Research and strategies for adapting formative assessments for students with special needs. In Andrade, H., & Cizek, G. J. (Eds.). (2010). *Handbook of formative assessment*. Routledge.

Furtak, E. & 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.

Furtak, E., Ruiz-Primo, M.A., Shemwell, J.T., Ayala, C., Brandon, P.R., Shavelson, R.J., & Yin, Y. (2008). On the fidelity of implementing embedded formative assessments and its relation to student learning. *Applied Measurement in Education, 21*(4), 360-389.

Gallagher, C. (2007). *Reclaiming assessment: A better alternative to the accountability agenda*. Portsmouth, NH: Heinemann.

Hattie, J. & Jaeger, R. (1998) Assessment and classroom learning: a deductive approach. *Assessment in Education*, *5*(1), 111–122.

Hattie, J. & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81-112.

Hambleton, R. K. , & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer.

Hambleton, R., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory*. London: Sage.

Hughes, G. B., Formative assessment practices that maximize learning for students at risk. In Andrade, H., & Cizek, G. J. (Eds.). (2010). *Handbook of formative assessment*. Routledge.

Jessup, L. M., Connolly, T., & Tansik, D. A.(1990). Toward a theory of automated group work: The de-individuating effects of anonymity. *Small Group Research*, *21*(3), 333–347.

Lucks, M. (2003). How do we get there from here? Formative assessment and feedback practices in two middle school science classrooms. Unpublished master's thesis, Stanford University, Stanford, CA.

Lu, R., & Bol, L. (2007). A comparison of anonymous versus identifiable e-peer review on college student writing performance and the extent of critical feedback. *Journal of Interactive Online Learning*, *6*(2), 100–115.

Lundstrom, K., & Baker, W. (2009). To give is better than to receive: The benefits of peer review to the reviewer's own writing. *The Journal of Second Language Writing,18*(1), 30–43.

Massey, F.J. (1951). The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association*, *46*(253), 68–78.

MacLeod, L. (1999). Computer-aided peer review of writing. *Business Communications Quarterly*, *62*(3), 87–95.

Mason, V. J., & Bruning, R. (2001). *Providing feedback in computer-based instruction: What the research tells us.* Lincoln, NB: Center for Instructional Innovation, University of Nebraska-Lincoln. Retrieved May 15, 2015, from http://dwb4.unl.edu/dwb/Research/MB/MasonBruning.html

Masters, G. N., & Wright, B. D. (1997). The partial credit model. In *Handbook of modern item response theory* (pp. 101-121). Springer New York.

McMillan, J. H., Hellsten, L., & Klinger, D. (2007). *Classroom assessment: Principles and practice for effective standards-based instruction* (4th ed.). Boston: Allyn & Bacon.

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

Moreno, R. (2004). Decreasing cognitive load for novice students: Effects of explana- tory versus corrective feedback in discovery-based multimedia. *Instructional Science, 32*, 99–113.

Nyquist, J.B. (2003). *The benefits of reconstructing feedback as a larger system of formative assessment: A meta-analysis*. Unpublished master's thesis. Nashville, TN, Vanderbilt University.

Narciss, S., & Huth, K. (2004). How to design informative tutoring feedback for multi-media learning. In H. M. Niegemann, R. Brunken, & D. Leuther (Eds.), *Instructional design for multimedia learning* (pp.181–196). Munster, NY: Waxmann.

National Research Council. (2011). *Expanding underrepresented minority participation: America's science and technology talent at the crossroads.* Committee on Underrepresented Groups and the Expansion of the Science and Engineering Workforce Pipeline. Washington, DC: The National Academies Press.

National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas.* Committee on Conceptual Framework for the New K-12 Science Education Standards. Board on Science Education. Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

National Research Council. (2014). *Developing Assessments for the Next Generation Science Standards.* Committee on Developing Assessments of Science Proficiency in K-12. Board on Testing and Assessment and Board on Science Education, J.W. Pellegrino, M.R. Wilson, J.A. Koenig, and A.S. Beatty, *Editors.* Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

NGSS Lead States (2013). *Next Generation Science Standards: For States, By States.* Washington, DC: The National Academies Press.

Nichols, S., & Berliner, D. (2007). *Collateral damage: How high-stakes testing corrupts America's schools.* Cambridge, MA: Harvard Education Press.

O'Donnell, A. M., & Topping, K. (1998). Peers assessing peers: Possibilities and problems. In K.J. Topping & S. Ehly (Eds.), *Peer-assisted learning*, 255-278 (pp. 255–278). Mahwah, NJ:Erlbaum.

Pellegrino, J.W., Wilson, M.R., Koenig, J.A., & Beatty, A.S. (Eds.). (2013). *Developing assessments for the Next Generation Science Standards.* Washington DC: National Academies Press.

Popham, W.J. (2008). *Transformative assessment.* Alexandria VA: Association of Supervision and Curriculum Development.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2002). Reliable estimation of generalized linear mixed models using adaptive quadrature. *The Stata Journal*,*2*(1), 1-21.

Ramsden, P. (2003). *Learning to teach in higher education*. London: Routledge.

Richer, D. L. (1992). The effects of two feedback systems on first year college students writing proficiency. Dissertation Abstracts International, 53(08A), 2722.

Robinson, J. (1999). Computer-assisted peer review. In Brown, S., Bull, J. and Race, P (Eds.), *Computer-Assisted Assessment in Higher Education* (pp.95–102). London: Kogan Page.

Ruiz-Primo, M.A., Furtak, E., Ayala, C., Yin, Y, Shavelson, R.J. (2010). Formative assessment, motivation, and science learning. In Andrade, H., & Cizek, G. J. (Eds.).(2010). *Handbook of formative assessment*. Routledge.

Sadler, D.R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*(2), 119–144.

Sadler, D.R. (1998). Formative assessment: Revisiting the territory. *Assessment in Education*, *5*(1), 77–84.

Schaffer, J. (1996). Peer response that works. *Journal of Teaching Writing*, *15*(1), 81–89.

Scriven, M. (1967). The methodology of evaluation. In R.W.Tyler, R.M. Gagne, &M.Scriven (Eds.), *Perspectives on curriculum evaluation* (pp.1-22). Mahwah, NJ: Erlbaum.

Shavelson, R. J. Stanford Educational Assessment Laboratory, & Curriculum Research and Development Group (2005). *Embedding assessments in the FAST curriculum: The romance between curriculum and assessment*. Stanford, CA: Authors.

Shavelson, R. J., Yin, Y., Furtak, E.M., Ruiz-Primo, M. A., &Ayala, C. (2008). On the role and impact of formative assessment on science inquiry teaching and learning. In J. Coffey, R. Douglas, & C. Stearns (Eds.), *Assessing science learning: Perspectives from research and practice* (pp. 21–36). Arlington, VA: National Science Teachers Association.

Shepard, L.A. (2006). Classroom assessment. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 624-646). Westport, CT: Praeger.

Shepard, L. A. (2008). Formative assessment: Caveat emptor. In C. Dwyer (ed.), *The Future of Assessment: Shaping Teaching and Learning* (pp. 279–303). NY: Lawrence Erlbaum Associates.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, *78*(1), 153–189.

Songer, N.B., Myers, P., & Beach, J. (2009) Change Thinking for Global Science: Fostering and evaluating inquiry thinking about the ecological impacts of climate change. Grant funded by the National Science Foundation.

Songer, N.B., Dewey, T., Peters, V., Reicher, M., Kwok, A., Zaidi, S., Tupper, B., Hammond, G., Myers, P. (2014) *Climate Change and Impacts on Ecosystems: A Middle School Curricular Unit.* Ann Arbor, MI: The University of Michigan.

Songer, N.B., Zaidi, S., Newstadt, M. (2016). Classroom-based holistic assessment and evidence of middle school students' Next Generation Science Knowledge. Manuscript in preparation.

Songer, N.B., Zaidi, S., Newstadt, M. (2016). Holistic assessment outcomes for middle schoolers' Next Generation Science Knowledge on climate change. Manuscript in preparation.

StataCorp. (2011). *Stata Statistical Software: Release 12*. College Station, TX: StataCorp LP.

Stemler, S.E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating inter-rater reliability. *Practical Assessment, Research & Evaluation*, *9* (4).

Stiggins, R., & Chappuis, J. (2005). Using student-involved classroom assessment to close achievement gaps. *Theory into practice*, *44*(1), 11-18.

Topping, K. J. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research*, *68*(3), 249–276.

Topping, K.J. (2010). Peers as a source of formative assessment. In Andrade, H., & Cizek, G. J. (Eds.). (2010). *Handbook of formative assessment*. Routledge.

Wiggins, G. (2006). Healthier testing made easy. Retrieved March, 2015, from http://www.edutopia.org/authentic-assessment-grant-wiggins

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

Wiliam, D., & Thompson, M. (2007) Integrating assessment with instruction: what will it take to make it work? In C. A. Dwyer (Ed.) *The future of assessment: shaping teaching and learning* (pp. 53-82*).* Mahwah, NJ: Lawrence Erlbaum Associates.

Yang, J. C., Ko, H. W., & Chung, I. L. (2005). Web-based Interactive Writing Environment: Development and Evaluation. *Educational Technology & Society*, *8*(2), 214–229.

Yin, Yue, Shavelson, Richard J., Ayala, Carlos C., Ruiz-Primo, Maria Araceli, Brandon, Paul R., Furtak, Erin Marie, Tomita, Miki K. and Young, Donald B. (2008). On the impact of formative assessment on student motivation, achievement, and conceptual change. *Applied Measurement in Education, 21*(4), 335- 359.

Zhao, H. (2014). Investigating teacher-supported peer assessment for EFL writing. *ELT Journal*, *68*(2). Doi:10.1093/elt/ccto68

Chapter 3

Using Anonymous Reviews of Peers and Expert to Support the Construction of *Next Generation Science Knowledge* of Middle School Students

Introduction

The new *Framework* for K-12 science education (NRC, 2012) advocates for promoting integrated knowledge in which *disciplinary core ideas* in the science and engineering disciplines are learned through the *practices* used by scientists and engineers (NRC, 2012). This shift is from a focus on one dimension in isolation, such as factual disciplinary core ideas to be memorized, to a focus on fostering the learning of disciplinary core ideas through the science and engineering practices (NRC, 2012; Reiser, 2013; Songer &Kali, 2014). Furthermore, the *Framework* proposes that efforts to foster integrated knowledge (hereafter referred to as "NGSK" (see Chapter 1, this dissertation)) should illuminate to students the activities undertaken by scientists and engineers in building and communicating knowledge (NRC, 2012). It is through this endeavor that the *Framework* intends to promote science learning in K-12 settings that bears connections to how science is practiced by scientists and engineers (NRC, 2012).

Research suggests that one way to support knowledge building to illuminate to students the activities of professionals in that discipline is through the use of apprenticeship models (Collins, Brown & Newman, 1989; Resnick 2010). Apprenticeship models challenge students to participate in the intellectual work of the discipline (see Chapter 1, this dissertation). By situating learning in purposeful and engaging activities learners can be apprenticed into the activities of professionals in the discipline along with learning disciplinary content and practices (see, Bransford et al, 1999; Collins, Brown & Newman, 1989).

The *Framework* (NRC, 2012) identifies that when building knowledge, scientists and engineers engage in the activity of receiving and providing critique on their knowledge products (NRC, 2012). Thus, "critique is an essential element both for building new knowledge in general and for the learning of science in particular" (NRC, 2012; p.43). Despite scientists and engineers engaging in the activity of receiving and providing critique there is minimal use of engaging in critique during knowledge building in K-12 science classrooms (NRC, 2012). Responding to the need to promote knowledge building that uses the activity of receiving and providing critique, this study employs the activity of *review* that scientists and engineers engage in to receive and provide critique on their work.

The activity of review is used within a classroom formative assessment system (the *4R Activity*, see Chapter 1, this dissertation) designed to support students in developing proficiency in NGSK (Songer et al., in preparation). The assessment system is embedded in a middle school climate change curriculum designed to foster NGSK (Songer et al., 2014) (see Chapter 1, this dissertation).

In the 4R Activity, the review activity functions as a feedback mechanism to provide personalized feedback to students targeted towards strengthening the NGSK product (see Chapter 1, this dissertation). Furthermore, the review activity is designed to simulate the double-blinded review process of scientific journals (see Chapter 1, this dissertation) and through it directly engages students in an activity that reflects how scientists and engineers receive and provide critique.

Research Questions

In this study student learning was investigated in three areas: (1) incorporating feedback to revise the NGSK products; (2) providing review; and (3) reflecting on how to improve in providing review. The following research questions guided the analysis:

1. With respect to the scores of the original NGSK products, what trends in scores are seen for the revised NGSK products constructed during the 4R Activity?
2. What patterns and characteristics are observed over time in the peer review provided by students in the 4R Activity?
3. What patterns and characteristics are observed over time in the reflection undertaken by students on their peer reviewing skills in the 4R Activity?

By examining the trends in scores received by students on their original and revised NGSK products, learning associated with incorporating feedback was investigated. Learning associated with providing feedback was explored by examining the patterns and characteristics of students’ peer reviews. Lastly, learning associated with how students reflect on ways to improve in reviewing was explored through the emerging patterns in student’s reflections on his/her reviewing skills.

Relevant Background Research

This study is informed by research in science education, on the practice of constructing evidence-based explanations, and research that discusses affordances and challenges of engaging students in the practice of peer review.

Constructing Evidence-based Explanations

In the 4R Activity as well as in the curriculum, the NGSK products were constructed as evidence-based explanations or predictions. The distinction made between an evidence-based explanation and an evidence-based prediction in the curriculum was that a prediction concerned explaining a phenomenon for a future event, such as the affect of climate change on species distribution in the future. This distinction is supported by research in the philosophy of science (see, Rescher, 1958; Hanson, 1959) that has emphasized the logical similarities between prediction and explanation with the point of departure between the two being the time of occurrence of the event for which the explanation or prediction is being constructed.

The practice of constructing evidence-based products is considered as “integral to the core work of science” (Windschitl, Thompson & Braaten, 2008; p. 3). Similar to the NGSK products in the curriculum, the NGSK products in the 4R Activity comprised a *claim* for which relevant evidence (*data*) needed to be provided along with *reasoning* to link the evidence to the claim. The three parts of the explanation were adapted by the research team from Toulmin’s (1958) model of argumentation to support the practice of constructing explanations in the science classroom, which have also been used extensively by other researchers (Bell & Linn, 2000; Jimenez-Aleizandre, Rodriguez, & Duschl, 2000; McNeill, Lizotte, Krajcik & Marx, 2006; Songer, Kelcy & Gotwals, 2009;

Gotwals & Songer, 2012). Additionally, the use of an adapted form of Toulmin's (1958) model of argumentation in the construction of NGSK products is consistent with the work of other science educators (Lee, 2003; Sandoval & Millwood, 2005; McNeill et al., 2006) that work with elementary, middle and younger high school students and focus on creation of integrated knowledge products (like NGSK) consisting of the practice of scientific explanations.

There is much research in the field of science education (see Songer & Gotwals, 2006; Songer, Kelcy & Gotwals, 2009; McNeill, Lizotte, Krajcik & Marx, 2006) that shows successful curricular interventions that foster students' construction of explanations (the NGSK products in this study used the science practice of constructing explanations). However these studies and others have also highlighted various difficulties associated with students' construction of scientific explanations. For example, Sadler's (2004) work documents the challenges associated with students' understanding of what counts as evidence, while Sandoval's (2003) work highlights the struggles associated with students' use of appropriate evidence and the difficulty of supporting their claims with sufficient evidence (Sandoval & Millwood, 2005). Additionally other researchers (Gotwals & Songer, 2009; Bell & Linn, 2000) have shown that students have difficulty providing reasoning for how the evidence supports the claim. Specifically, middle school students have the most difficulty with the reasoning component of scientific explanations (Lizotte, Harris, McNeill, Marx & Krajcik, 2003; McNeill et al., 2006). With respect to middle school students, researchers (e.g., Gotwals & Songer, 2009) have discussed that many students possess "intermediary or middle knowledge" (Gotwals & Songer, 2009; p.276) pertaining to the construction of scientific explanations where parts of their explanation are correct but either parts of the disciplinary idea or the structure of the explanation is only partially accurate (Gotwals & Songer, 2009). These studies show that because of these challenges students need to be explicitly supported in the building of scientific knowledge through the construction of evidence-based explanations.

Researchers have proposed many strategies to support students, especially middle school students in the construction of scientific explanations (e.g., Songer & Gotwals, 2012; McNeill & Krajcik, 2012; Sandoval & Millwood, 2005). For example, Sandoval & Millwood (2005) propose the involvement of students in sustained epistemic discourse on their constructed explanations to support the development of disciplinary knowledge and the practice of constructing explanations. Songer and Gotwals (2012) discuss the continued use of scaffold-rich curricular activities to support students in constructing scientific explanations. Furthermore, McNeill and Krajcik (2012) propose various strategies for teachers to help them incorporate scientific explanation in their classrooms. This study proposes another strategy, that of using a feedback system to foster the construction of evidence-based scientific explanations (referred to as NGSK products).

In the study, active participation of peers and experts is utilized in the feedback process to support students' in constructing NGSK products. The 4R Activity functions as a periodic activity that is used by teachers to support students in building NGSK. As an activity embedded within the curriculum, the 4R Activity provides scaffolds in the form of built-in supports to students during the construction NGSK products, during peer feedback, and during the revision of the NGSK products. Additionally the feedback from peers and experts provides additional support to learners while functioning as a method to engage in discourse on the constructed scientific explanation.

Engaging Students in the Practice of Receiving and Providing Reviews

Engaging in the practice of reviewing (i.e., peer feedback) has a long history of success within educational research. Studies suggest that critical evaluation of a peer's work can foster self-assessment of one's work and enhance learning of both the parties that are engaged in receiving and providing review (Lunstrom & Baker, 2009; Liu &Carless, 2012). Feedback allows students to take on an active role in their own learning by providing them an opportunity to monitor their learning (Butler & Winnie, 1995;Venables & Summit, 2003). In addition to a teacher (or expert) providing feedback, peers can also be used as a source of providing feedback. Peer feedback is shown to promote learning for the student who receives feedback as well as the student who provides the feedback. In spite of the benefits of peer feedback there are only a handful of studies using written peer feedback for K-12 students. Additionally these studies use the peer feedback systems in disciplines other than science (e.g., Tseng & Tsai, 2007; Lu & Law, 2012).

However, research on the use of written peer feedback has been more extensively used in higher education settings. For example, research by Lundstrom and Baker (2009) and DeGuerro and Villamil (2000) with undergraduate students suggests that reviewers benefit from reviewing their peers' work as both the author and the reviewer learn from the peer review exercise. The author learns how to improve the reviewed product after receiving feedback while the reviewer can build competency in critically evaluating the work of the author, which might then aid in his/her own thinking and writing. Even when both the writer and the reviewer are novice learners, they scaffold each other's learning and benefit from receiving and providing feedback (Teo, 2006). Additionally, Shaw's

(2002) work suggested that students tend to write more carefully when they are aware that the audience for their writing is wider than just their teacher.

Even though there are many potential benefits of peer feedback, there are also challenges associated with peer feedback, such as bias in feedback arising out of stereotypes associated with gender, race or minority status. Cohen, Ross and Steele (1999) argue that feedback may be perceived as threatening for students from underrepresented groups. They argue that students from underrepresented groups might feel that they are being judged in light of a negative stereotype pertaining to the intellectual ability of their group rather than being judged on their own intellectual merits. This in turn might hinder the uptake of feedback and result in students from withdrawing from doing the task. This study addresses the problem by introducing a double- blinded feedback system, where the identity of the author and the reviewer are concealed from each other. Anonymity achieved through a double-blinded review process alleviates the bias caused by interpersonal dynamics and stereotypes associated with certain groups of students.

Research on blinding within the professional science community (Budden et al., 2008) found that the use of a double-blinded review process in an Evolutionary Biology journal lead to a significant increase in the representation of first-authored papers by females. This pattern was not observed in a similar journal (as claimed in the study) that did not adopt a blinded review process. Additionally, the study did not identify any negative effects due to blinded reviews. Within the educational setting, Lu and Bol's (2007) study comprising 92 undergraduate students in an English class revealed that those who received anonymous reviews from their peers outperformed their counterparts that did not receive anonymous review on their writing performance task. Additionally, the anonymous group provided more critical feedback. Web-based peer review systems that maintain anonymity during the review process also allow for the monitoring of progress of the reviewers and authors in real time (Lin, Liu & Yuan, 2001). Moreover, this study argues that by using web-based blinded review processes similar to those used in academia, students can be apprenticed into the present day peer review practices of scientists and engineers.

Supporting Students in the Practice of Providing Reviews

Unlike professional scientists and engineers, students are not trained in the process of providing peer reviews. Students may be starting to deeply learn the core concepts and disciplinary practices (e.g., constructing NGSK products) they are reviewing, and thus find providing critique challenging. These limitations might hinder the generation of effective peer reviews. For example, research that compared novice and expert's peer review practices highlighted that novices were more focused on style issues of writing rather than on content aspects of writing (Flower, Hayes, Carey, Schriver &Stratman, 1986). Furthermore, Bridwell's (1980) work showed that when providing peer review, 89% of high school students suggested sentence level changes such as grammar, spelling and punctuation issues. This study proposes that in order to support students in the development of their peer review skills and to provide a comprehensive peer review to the author, the use of an expert's review could be added. An expert's review could serve two purposes. First, it could build on the peer review and also provide review on aspects that were not covered by the peer. This would allow the author to receive a more comprehensive review that might aid in the revision process. The second benefit might arise for the reviewer. By having access to the review done by an expert a model of expert strategies are made available to the learner (cf. Collins & Brown, 1988). This is also supported by the work of Williams (1993) who argues that without modeling expert strategies to learners, the novice learner is left to acquire the strategies through trial and error, which might result in frustration for the learner and also result in the adoption of ineffective strategies.

Access to an expert's review is employed in the 4R Activity through a guided self-reflection. In the reflection activity learners revisit the peer review they provided in light of an expert's review. This allows learners the opportunity to draw comparisons and become sensitized to the differences in their own performance in relation to the expert's performance (Collins & Brown, 1988). By facilitating a comparison between a student's review and that of an expert, the guided self-reflection in the 4R Activity aimed at helping students become aware of reviewing strategies of an expert that they could then use in future reviewing activities.

Methods

This section provides information on the participants, the data sources, and the analysis employed to examine the data.

Participants

The student participants were situated across four partner school sites and in the classrooms of five teachers participating in the climate change curriculum project. All schools were located in the U.S.A. and were publically funded. Participant selection for this study was based on two criteria. First, students needed to have parental/guardian consent and their own assent for their information to be included in this study. Second, students had to have participated in all the three 4R Activities (see Chapter 1, this dissertation). This resulted in a sample size of 173 students in grades 6-8 across the four school sites.

In an effort to promote an inclusive learning environment with equitable learning opportunities, differentiated support was offered by the teachers to students of special needs and students classified as English Language Learners (ELLs) in the 4R Activity. Research on ELLs suggests that students classified as ELLs face a dual challenge: the challenge of English language acquisition and the challenge of learning disciplinary content in the English language that they are still acquiring (e.g., Abedi, 2010; Meskill, 2010). Given the linguistic complexity of the activity, the participating teachers supported ELL students based on the individual students needs. Similarly students with special needs (students with IEPs) were also offered supports during the 4R Activity to meet their learning needs (Elliot, Ryan, Kettler, Beddow &Kurz, 2010). Each teacher decided the type of support he/she would offer to students with IEPs. The author of this paper did not have authorization to inquire about the specific IEPs of students and therefore did not have information on the type of support received by students with IEPs beyond what the author observed in classroom observations. Table 3.1 presents the information on the participants, the school sites, and the five classrooms of the participating teachers of this study.

Table 3.1
Participant Information and school location

School ID	U.S. Census Bureau Designated Divisions	% Enrollment in Free/Reduced Lunch Program	Participating Teacher*	Teacher's Science Teaching experience (Years)	Grade	Participating Students	Students with Individualized Education Programs (IEPs)	Students designated as English Language Learners (ELLs)
B	New England	36	Ms. Lata	2	6	17	1	7
B	New England	36	Ms. Asha	6	6	17	1	8
C	West North Central	30	Ms. Priya	34	7	33	6	0
D	East North Central	75	Ms. Laila	18	7	30	4	0
E	East North Central	12	Mr. Amar	10	6,7,8	76	4	6
					Total	173	16	21

*Pseudonyms are used for the participating teacher

Data Sources

The primary data sources used in this study pertained to students written products that were generated in each stage (respond, review, revise, reflect) of the 4R Activity. The written data from the stages of the 4R Activity are presented in detail in the sections below. They are organized by the research questions that were being addressed by that data source. The three research questions are restated below:

1. With respect to the scores of the original NGSK products, what trends in scores are seen for the revised NGSK products constructed during the 4R Activity?

2. What patterns and characteristics are observed over time in the peer review provided by students in the 4R Activity?

3. What patterns and characteristics are observed over time in the reflection undertaken by students on their peer reviewing skills in the 4R Activity?

Additionally three other data sources were used to provide a deeper understanding for the patterns emerging from the students written products. These data sources were, (1) discussions with the participating teacher prior, during and after the implementation of the 4R Activity; and (2) field notes from classroom observations (6 observations per classroom) for each 4R Activity, (3) online log of submission that provided a timestamp for the submission.

Trends in the scores of constructed and revised NGSK products. Students NGSK products from stage 1 (respond) and stage 3 (revise) of the 4R Activity served as the data source used in determining the trends in score in the revised NGSK products with respect to the scores of the original NGSK products. Three sets of scores (respond-revise) pertaining to each 4R Activity were used as the data source. The data for the construction and revision of the NGSK products consisted of student scores on the three NGSK products. Both the constructed and revised NGSK products were scored. The scores on the NGSK products were not provided to the students. This decision was based on research (e.g., Kluger and DeNisi, 1996; Siero & Van Oudenhoven, 1995) that indicated negative effects on performance when grades or numeric scores accompanied feedback. The scores provided an evaluation to the learner and deviated the learners from incorporating the feedback.

The rubric used to score the NGSK products had been validated using think aloud interviews with students and through Item Response Modeling (Songer, Zaidi & Newstadt, in preparation). The rubric was designed on the criteria of scientific accuracy of information along with the measures of quality such as elaboration on the scientific information used and incorporation of scientific terminology The score range of the rubric was from Level 0 to level 5, with level 5 being the highest achievable level (refer to Appendix B for the rubric and the characteristics of each level).

In this study, if students achieved a level 4 or 5 on their NGSK products they were designated as proficient. This was because both levels 4 and 5 represented a complete NGSK product with the only difference between these levels being that of effective incorporation of scientific terminology in a level 5 product. Given the 4R Activity was administered to promote students' learning (a formative assessment) and not as an end of curriculum summative assessment, it was important to consider level 4 as a level that also demonstrated a level of proficiency.

Patterns and characteristics observed in students' reviews. Student reviews from the review activities were used as the primary data for the analysis of characteristics and patterns in student reviews. Across the three 4R Activities there were a total of 928 reviews consisting of both primary and additional area of improvements. As a secondary source of data, five scientist reviews from each class were also sampled to look at the similarities (if any) between the students and the scientist's reviews. In total there were 25 scientist reviews that were considered. These reviews were sampled from the first and second 4R Activities and were selected based on the representativeness of the review provided for that class. Representativeness was determined based on the frequency of two most commonly occurring types of review with respect to the selected area of improvement and suggestion provided by the scientist for a particular classroom. For example, in the first 4R Activity, the two most frequent reviews given by the scientist for Ms. Laila's class was the selection of Area C (need for specific reasoning) with an accompanying suggestion that reminded students what comprised as reasoning and provided hints for what could be included as reasoning in that particular NGSK product. Area D (scientific terminology), with suggestions of words/phrases that could be added to different parts of the written NGSK product, was the second most frequent type of review that Ms. Laila's class received in the first 4R Activity. Thus, the two scientist's reviews selected from Ms. Laila's class from the first 4R Activity were two reviews that focused on Areas C and D.

Reflecting on peer reviewing skills. Reflection was the last stage of the 4R Activity (see Chapter 1, this dissertation). The written reflections of students from the three 4R Activities were used as data for this analysis. In total there were 519 reflections from the three 4R Activities.

Analysis

The analysis pertaining to each of the three research questions is presented separately under different sections below. In each section, the research question being addressed is provided followed by the analysis employed to address the research question.

Trends in the Scores of Constructed and Revised NGSK Products

As a reminder to readers, the first research question was: *With respect to the scores of the original NGSK products, what trends in scores are seen for the revised NGSK products constructed during the 4R Activity?* In investigating this research question, the analysis proceeded by scoring of the original and the revised NGSK products using the rubric discussed previously (see Appendix B for the rubric). Reliability measures were conducted using the consensus estimates method (Stemler, 2004). In accordance with the consensus estimates methods, another rater was trained in how to interpret the scoring criteria outlined by the holistic rubric. This was achieved through multiple meetings between the author of this paper and the rater. In the meetings the criteria for scoring were discussed along with a demonstration of the scoring process. Additionally, the rater was provided with exemplar responses for each score category. Once the rater had been trained, the author of this paper and the rater scored 10 identical NGSK products that were randomly sampled from the data. This process was repeated until a 90% agreement was established for each of the three NGSK products. After a 90% agreement had been reached, the author of this paper proceeded with scoring the entire data. Spot checks on 10% of the data were performed after the scoring had been completed.

Once the scoring had been done, the percentage of students who showed at least a one score level increase in their revised NGSK product was determined. This was done for all participating classes and for all the three 4R Activities. This analysis provided information on the percentage of students who demonstrated any improvement on their revised NGSK product.

The second type of trend determined the proficiency levels (level 4 or 5) achieved by students on the revised NGSK product. For this analysis, the score level on the constructed NGSK product (referred to as the start level) was determined for those students who had attained proficiency levels (levels 4 and 5) on their revised NGSK product. The movement to proficiency level analysis was focused on investigating the start level of students who ended up at proficiency levels (level 4 or 5) on the revised NGSK product. This analysis provided insight on the most frequent type of movement in score levels that was observed for each class in each 4R Activity.

Patterns and Characteristics Observed in Students' Reviews

The second research question was: *What patterns and characteristics are observed over time in the peer review provided by students in the 4R Activity?* The analysis employed to investigate this research question was an inductive approach, where the coding of the reviews was done first by open coding the responses, then employing focused coding by development of categories (Maxwell, 2013) that emerged out of the data. Additionally deductive reasoning (Patton, 2002) was also employed when generating categories, which were based on existing literature on important dimensions of quality in a peer review (e.g., Shattel, Chinn, Thomas & Cowling, 2010; Harding, 2010). Examples of some of the open codes developed along with the inductive and deductive categories are shown in Table 3.2.

Table 3.2

Examples of a deductive category and the associated inductive categories, and focused codes in peer review data

Deductive Category	"Specific Suggestions" (Shattel, Chinn, Thomas & Cowling, 2010)	
Category Created Inductively	**Focused Coding**	**Data Segments** (Verbatim excerpts of student work)
Explicitly identifying the location of where to make the improvement in the NGSK product	Paraphrasing	"In your evidence you stated that Montana will be too cold, you did not state how cold it will be."
	Direct Quotes	"When you said "on the graph it shows the line going up," I don't know what line you are talking about, or even what graph you are talking about."
Specific Suggestion given on how to make improvements	Example	"In the reasoning, you need to talk about the predator-prey relationship for fox and the shrew."
	Hints	"You could use another source to really make the evidence solid, (hint: the simulations we did in lesson 5 that shows what happens to sun's energy when GHGs increase. That can be a good source of additional evidence)…"
	Sentence Replacement	"In your evidence, include where you saw that the brown squirrel can live in Canada. Explain what you specifically saw that made you conclude on what you did. You could add something like: "In the future, the brown squirrel will live in Canada because on the prediction map Canada was the highlighted region." Somewhere along those lines..."

Student Reflections on the Practice of Reviewing

The analysis on student reflection on the reviewing skills was guided by the third research question: *What patterns and characteristics are observed over time in the reflection undertaken by students on their peer reviewing skills in the 4R Activity?* Similar to the analysis employed for the second research question about students' reviews, the data of students' reflections were open coded and then focused coding was employed. This was followed by formation of categories that resulted in characteristics of reflection grounded in the data (Emerson, Fretz & Shaw, 2011). Deductive codes from research were not employed due to the highly contextual nature of the reflection activity. An example of some of the open codes along with the inductively generated category is shown in Table 3.3.

Table 3.3

Examples of an overarching category, and the associated focused and open codes from data on student reflections

Category	Focused Codes	Open Codes	Data Segment (direct quotes)
Reflection stemming directly from the scientist's feedback to the reviewer	**Verbatim:** Includes all aspects of what the scientist said (e.g., directly quoting the scientist's feedback, or making minor changes to the scientist's feedback)	Minor change (e.g., changing the person pronoun from the scientist's feedback.	"What i learned is to provide examples of what sentences need to be corrected."
		Direct quote from the scientist's feedback	"In the future, try to add example sentences or words that they can add"
	Paraphrased from scientist's feedback	Summarization of the scientist's feedback.	"One thing that I learned from reading the scientist's review is that I could tell them [the student's whose work was being reviewed] the reason for making the changes that I suggested. This will help them understand why my changes will make their prediction better."
	Elaborated reflection on only the aspects that were present in the scientist's feedback to the reviewer.	Details added to support the scientist's feedback to the reviewer.	"I need to describe more when I'm reviewing because I kind of just say you need to fix "that" but I think I need to go into more detailed and say you need to fix what your saying about 1900-2010 because my peer might not understand what I'm saying in the review."

Results

The results pertaining to: (1) revising NGSK products; (2) reviews done by students; and (3) reflections on the peer reviews are presented in this section. In the 4R Activity, students received reviews that they used in revising their NGSK product, provided peer review, as well as engaged in a reflection aimed at improving peer review skills.

1. Trends in the Scores of Original and Revised NGSK Products

The results on the revised NGSK products show that over time there was an increase in the score levels on the revised NGSK product. Additionally students who began at lower start levels (e.g. Level 0, 1 or 2) were also able to reach proficiency levels (Level 4, 5) on their revised NGSK products. These results are presented in the following two sections. The first section presents percentage distribution of students who demonstrated an increase of at least one score level on their revised NGSK product when compared to the score level of their original NGSK product along with the distribution of students who did not show score improvements. The second section showcases the start levels of those students who achieved a proficiency score level (level 4 or 5) on their revised NGSK products. With one of the goals of the 4R Activity being to promote student learning of NGSK, it was important to track the students who achieved proficiency levels by the end of the activity. Results pertaining to the above two trends allow for the determination of two salient points. First, the results on the distribution of students based on score improvement provide an overview of whether the revision process was leading to an improved NGSK product. Second, the results from the proficiency level analysis provides information on the different start levels at which students began who were able to reach proficiency levels at the end of the 4R Activity.

a. Score Improvements between the Original and the Revised NGSK Products

The improvement in score between the original and revised product was defined as an increase in at least one score level from the original to the revised NGSK product. The score improvements are presented as percentages in Figure 3.1. Students who had already achieved a score level of 5 on their original NGSK product could not be represented under the score improvement category and were represented under the category “start score level of 5”. It is to be noted that the results pertain to only score improvement and therefore some students with original NGSK products at a level 5 made improvements to their NGSK product during revision but this improvement could not be captured in their scores as they had already attained the highest score on their original NGSK product. For example, students at a start level of 5 made improvements by including more scientific terminology, fixing spelling and punctuation errors, and in some cases added additional evidence to their NGSK product from the lessons in the curriculum. For the students who did not show a score improvement on the revised NGSK product and had not already attained a level 5 on their original NGSK product, their original and revised NGSK products were compared to determine if they had made any changes during revision. The students who made some revisions to their original NGSK product but their revisions were not substantial enough to have lead to a score increase were represented under the category “minimal changes made during revision”. Minimal changes during revision were categorized as implementing only feedback on writing aspects such as correcting grammar, correcting punctuation, capitalizing appropriate words, and fixing spelling errors when more substantial changes (e.g., adding specific evidence) had also been suggested. The last category of students who did not show a score improvement were those who had made no changes to their NGSK product. They were represented in the category "no changes made during revision".

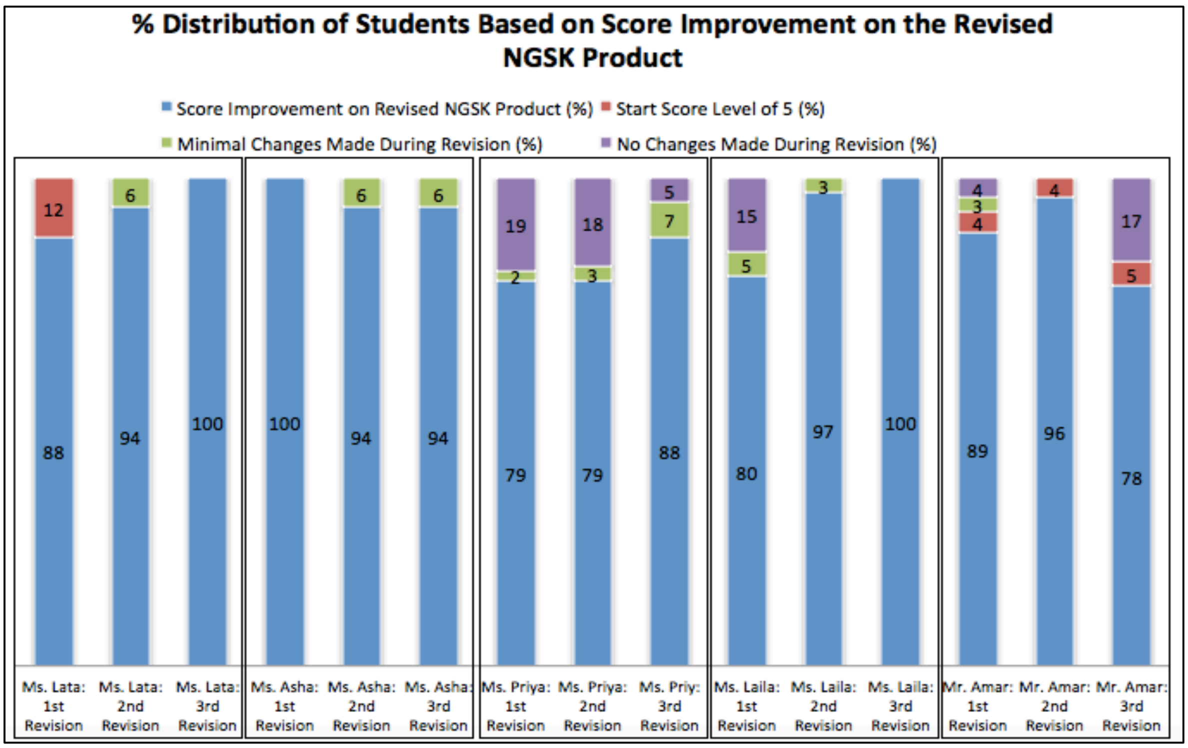


Figure 3.1. Percent distribution based on score improvement on the revised NGSK product.

In Figure 3.1, readers will see that in successive revision activities a higher percentage of students show a score improvement on the revised NGSK product. Additionally by the last revision activity, there was a decline in the percentage of students who did not make changes to their original NGSK product. However the last revision activity of Mr. Amar's class was an exception to the above two trends. The 17% of students who did not show a score improvement on the revised NGSK product were all from one section of Mr. Amar's class. In this section, a substitute teacher had administered the revision activity as Mr. Amar was on a leave of absence. Classroom observations and the time stamp for the submissions revealed that the students submitted their revised responses much more quickly than the usual time they took to submit their responses in the previous revision activities (when Mr. Amar was present). The presence of a substitute teacher could be one of the reasons for students not making changes during revision and therefore their revised NGSK product remaining at the same score level as their original NGSK product.

The percentage of students who did not make any changes during the first revision activity was high in Ms. Priya and Ms. Laila's classes. In both these classes, classroom observations and discussion with the teachers revealed that students were hesitant to make revisions because they expressed that they were satisfied with their original NGSK product. Discussions with Ms. Priya and Ms. Laila also brought to light two possible reasons for the lack of revisions being made. The first was that students in both of these classes had expressed that they felt that making revisions implied that their original work was not satisfactory and they did not agree with this evaluation. It was only after the end of the first 4R activity that the teachers re-emphasized how revisions were an important part of learning and that having to make revisions only implied that the original NGSK product could be made stronger. The second reason provided by the teachers was that students were learning how to effectively incorporate feedback during the revision process and were therefore unable to do so in the earlier revision activities. Similar discussions with Mr. Amar on students making only minimal changes during revision made apparent that a reason for making minimal changes was that students were not practiced at making revisions to their responses, particularly in the science classrooms.

There were a small proportion of students across the dataset that had been classified as English Language Learners (ELLs), but Ms. Lata and Ms. Asha's classes had a very high proportion of students classified as ELLs. In their classes, roughly 50% of the students were classified as ELLs. In all the 4R Activities, the ELL students in Ms. Lata and Ms. Asha's classes showed score improvement. Both Ms. Lata and Ms. Asha worked with the ELL students during construction of the NGSK product and during the revision. The decision to offer support to students in all the classrooms was made by the teacher depending on his/her knowledge of the type of support required by the students. In discussions with Ms. Asha and Ms. Lata, the teachers stated that given the high language demands of the activity the students classified as ELLs in their class should be supported in constructing the NGSK product as well as in interpreting the feedback. The support took the form of reading the activity prompts, reading the feedback to the students and providing help with the vocabulary to aid in interpretation of the feedback.

Similar to Ms. Lata and Ms. Asha, Mr. Amar also worked with the ELLs during the construction of the NGSK product as well as during revision. He also offered supports similar to those offered by Ms. Lata and Ms. Asha during the construction of the NGSK product and during the revision process. In Mr. Amar's class there were 8% of students classified as ELLs. Out of these 7% showed a score improvement on at least two 4R Activities while 1% had already attained a level 5 on the original NGSK product.

With respect to the score improvements of students with IEPs, students in all classes demonstrated score improvement. Similar to the support offered to ELL students, Ms. Lata and Ms. Asha also supported students in Individualized Education Programs (IEP). There was only one student in Ms. Lata and Ms. Asha's class respectively with an IEP and they supported the student during the revision process. Mr. Amar had 4 students (5% of the class) with an IEP who were supported by an assistant teacher in both the construction and the revision of the NGSK product. Mr. Amar only checked in with the students with IEPs by inquiring about how far along they were with the construction of NGSK products and during the revision process. The author of this paper did not have information available on what the IEPs were of the students and was not authorized to ask the teacher specifically for that information. Perhaps his decision for the type of support offered to the students with IEPs was informed by the type of IEP that the different students had.

Ms. Laila and Ms. Priya's classes had four (13%) and six (18%) students with IEPs respectively. The students with IEPs worked independently in these two classes with occasional check-ins by the teacher during the construction and revision of the NGSK products. In both these classes also, all IEP students demonstrated score improvements.

b. Exploring the Start Levels of Students Who Ended Up at Proficiency Levels

The results discussed in this section explore the different start levels at which students began who attained proficiency levels. Figure 3.2 represents the start levels for each class along with the percentage of students from that class that began at that particular start level and attained level 4 or 5 in the revise stage of the 4R Activity. In Figure 3.2, the x-axis corresponds to the start levels of students who attained a level 4 or 5 respectively. This information is presented for each teacher's class for the three revision activities. The number on top of the bar shows the percentage of students from the total number of students in the class that went from a particular start level and attained a level 4 or 5 respectively. Summing the percentages for a given class would provide the percentage of students in that class that achieved a proficiency level of 4 or 5. Additionally, summing the percentages of students in a class that attained a level 4 or 5 would provide the total number of students who attained a proficiency level on the revised NGSK product.

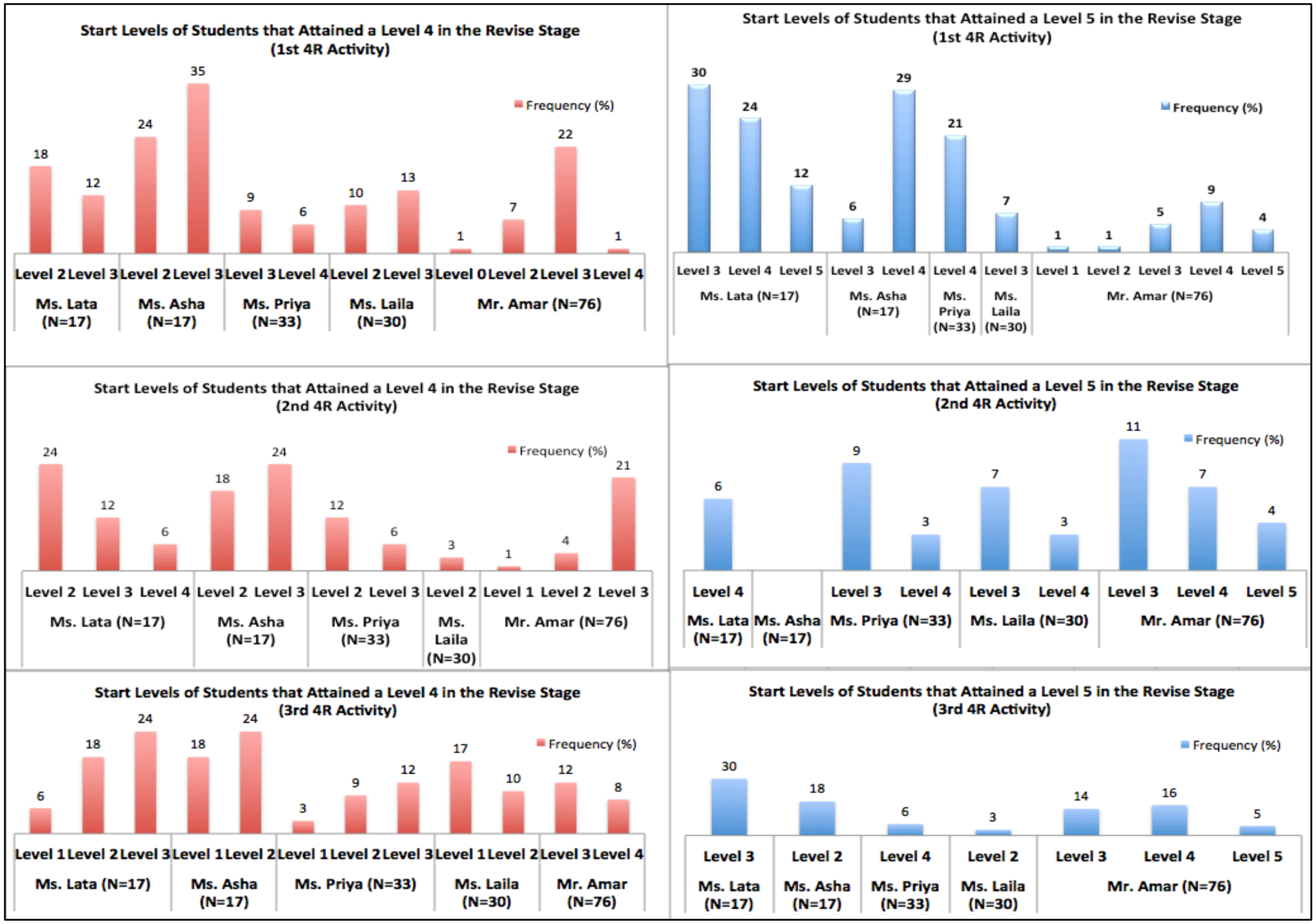


Figure 3.2. Start levels of students who achieved proficiency levels on the revised NGSK product.

In the 1st 4R Activity, students whose revised NGSK product was at a score level of 4 had most frequently attained a start level of 3. However, students whose NGSK products had a lower start level particularly level 2 (e.g., Ms. Lata's class, Ms. Laila's class) were also seen attaining a level 4 on their revised NGSK product. Additionally, it was observed that some students in Ms. Priya and Mr. Amar's classes that had achieved a start level of 4 remained at that level on their revised NGSK product. One possible reason for this was the resistance of students to revising their NGSK product as was observed and reported by teachers. This was particularly prevalent for students in Mr. Amar's class that had achieved a high start level (e.g., level 4). As discussed in the previous section, students had expressed that revision implied that their original NGSK product was not strong and they did not agree with this evaluation. Some of the students refused to make any revisions while others made some revisions, which were not substantial enough to improve the score level of their revised NGSK product.

When determining the start levels of students whose revised NGSK product was a level 5 in the first 4R Activity, the most frequently observed start level was a level 3 or 4. Also in Mr. Amar's class a very small percentage (1%) of students began at level 1 or 2 and attained a level 5. As seen in Figure 5, there were some students in Ms. Lata and Mr. Amar's classes whose original NGSK products were already at level 5 and these students remained at that score level. As previously discussed, students who began at level 5 also made revisions. However there revisions could not be reflected in a score level increase.

In the second 4R Activity, a level 4 on the revised NGSK product was most frequently seen for students who had a start level of 2. Additionally, students with a start level of 3 were also frequently seen as achieving a level 4 on the revised NGSK product.

As previously discussed, constructing the NGSK product in the 2nd 4R Activity was the most difficult of all the NGSK products constructed in the 4R Activities. This could be a factor in seeing a decline from the first 4R Activity in the percentage of students achieving proficiency levels. However, in Ms. Laila and Mr. Amar's classes, compared to the first 4R Activity there was a slight increase in the percentage of students who attained a level 5 on their revised NGSK product.

Overall, attaining a level 5 on the revised NGSK product was most frequently observed for students whose original NGSK products were levels 3 or 4. This was similar to that observed in the first 4R Activity. As can be seen in figure 5, no student from Ms. Asha's class achieved a level 5 in the second 4R Activity, however many students from her class achieved a level 4.

In the third 4R Activity, level 3 was the most frequent start level corresponding to a level 4 on the revised NGSK. This was similar to that observed in the previous 4R Activities. In addition to students whose original NGSK product was a level 2, students at a level 1 were also frequently seen attaining a level 4 on the revised NGSK product. Attaining a level 5 on the revised NGSK product was most frequently observed for students with a start level 3 or 4. This was similar to the observed trends in the previous 4R Activities. However, there were also students with a start level 2 (e.g., Ms. Asha and Ms. Laila's classes) who had achieved a level 5 on their revised NGSK product in the third 4R Activity.

The above trends included students with IEPs and students classified as ELLs. Nonetheless the trends of ELL students and students with IEPs are discussed below to highlight the proficiency levels attained by them. In Ms. Lata and Ms. Asha's classes, the one student with an IEP was able to reach proficiency levels in all the 4R activities. In Ms. Priya's class one out of the six students with IEPs reached proficiency levels on all the three 4R Activities. In Ms. Laila's class, two out of the four students with IEPs were able to reach proficiency levels on two out of the three 4R Activities. In Mr. Amar's class out of the four students with IEPs, no student reached proficiency levels on any of the three 4R Activities.

Students classified as ELLs were only in Ms. Lata, Ms. Asha and Mr. Amar's classes. In Ms. Lata's class six out of the seven ELL students reached proficiency levels on all three 4R Activities, while seven out of the eight students reached proficiency levels on all three 4R Activities in Ms. Asha's class. In Mr. Amar's class three out of the six students reached proficiency levels on all 4R Activities, while two of them attained proficiency levels on the first 4R activity only.

Summarizing the trends across the three 4R Activities it was seen that with each successive 4R Activity, students from start levels lower than 3 were also able to reach a proficiency level of 4 by the end of the 4R Activity (e.g., 60% of students in Ms. Asha's class that began at level 1 or 2 reached proficiency levels). Attaining a level of 5 by the end of the 4R Activity continued to be achieved by students who began at a level 3 or 4 (e.g., see Mr. Amar's class in Figure 5). However, in the last 4R Activity, students from a start level of 2 were also seen achieving a level 5 by the end of the activity (e.g., see Ms. Laila's class in Figure 5). Students with IEPs and students classified as ELL were also able to reach proficiency levels in the 4R Activity (e.g., Ms. Lata and Ms. Asha's class had approximately 86% of IEP students reach proficiency levels). These data show that a large proportion of ELL students reached proficiency levels while students with IEPs who reached proficiency levels were fewer. It is important to note that the sample size for students who are classified as ELLs was very small with the number of students with IEPs being even smaller. Additionally a large portion of the students with IEPs were from Ms. Laila and Ms. Priya's class where these students worked independently with only minimal support from their teacher. Even though Ms. Lata and Ms. Asha's classes had only one student with an IEP, in both these classes the student could reach proficiency levels on all the three revised NGSK products, perhaps because of the continued support of the teacher in helping the student during the construction and revision of the NGSK product.

2. Patterns and Characteristics of Students' Reviews

The emerging patterns and characteristics from students' reviews showed that with time students provided more specific and elaborated feedback as reviewers. The results pertaining to the review done by students in the 4R Activity are presented below. The first set of results pertains to the patterns observed in the areas of improvement (primary as well as additional) that the students chose in their review. The trends of students classified as ELLs is also discussed separately to provide insight into the review choices made by ELL students. There was no consistent trend observed in selection of the areas of improvement that was observed for students with IEPs within a class or across the classrooms. Therefore the trends of students with IEPs are not summarized separately but are a part of the overall trends discussed.

The second set of results describes the characteristics of the suggestions for improvement provided by students in the review process. In this, the characteristics are drawn from all the areas of improvement that were selected by the students. Since there was no underlying trend observed within the ELL students or students with IEPs, the characteristics of their reviews are not discussed separately.

a. Patterns in the *Area of Improvements* Students Chose to Review

When looking at patterns of what the students chose to review, it was observed that in the first review activity, 28% of students across the classes chose "Other Improvement" (Area E) as the primary area of improvement and 41% did not choose an additional area of improvement. By the third review activity, only 12% of students chose Area E as the primary area of improvement while the percentage of students who did not provide an additional area of improvement dropped to 14% by the third review activity. Out of the 12% of students who chose Area E as the primary area of improvement in the third review activity, 10% of them were students classified as ELLs.

By the 3rd 4R Activity, 53% of students listed Area E as an additional area instead of a primary area of improvement. This shows that when students first began to review, they were more focused on the writing aspects (e.g., grammar, punctuation, spelling errors) of the NGSK product and used it as the primary area of review. The high frequency of Area E as the most reviewed area by students in the first 4R Activity is consistent with research on student engaging in peer review that shows that students choose to review writing aspects of peers' written work as opposed to aspects pertaining to the content of the work (e.g., Flower, Hayes, Carey, Schriver &Stratman, 1986; Bridwell, 1980). The shift in Area E from a primary area to an additional area of improvement suggests that students began to look for improvements pertaining to the content of the NGSK product.

In the three review activities the focus on the inclusion of scientific terminology (Area D) increased over time as an additional area of improvement and declined as a primary area of improvement. Area D as a primary area of improvement declined from 19% to 7% over the three 4R Activities while it rose from 19% in the first review activity as an additional area of improvement rose to 60% in the last review activity. 95% of students classified as ELL chose Area D as an additional area of improvement.

One reason for the increase in Area D as an additional area of improvement could be that across all the schools, there was a growing emphasis placed by teachers on the inclusion of scientific terminology in the construction of NGSK products. This was also in line with the emphasis placed on the inclusion of scientific words in the curriculum itself. Additionally, the scientist's feedback to the reviewers that did not select an additional area of improvement often included the consideration of Area D as an additional area to consider when reviewing in the future. Thus a multitude of factors could have contributed to the increase of Area D as an additional area of improvement. As for the decline in Area D a primary area of improvement, it could be due to its increase in the category of "additional area of improvement".

As for the other areas of improvement, the percentage of Area A (missing parts in the NGSK product) as a primary area of improvement declined from 10% to 7% over time. The decline in Area A is understandable because over time students NGSK products that had missing parts also declined. With this decline the applicability of Area A in a review also declined. However Area B and C increased over time. Area B (need specific evidence) rose from 23% to 39% and Area C (need specific reasoning) rose from 20% to 35% across the three 4R Activities.

From the above patterns, it can be seen that in the first review activity, the percentage of students choosing Area B, C, and D was each around 20% while Area E was 28%. By the third review activity, Area B (need specific evidence) and C (need specific reasoning) became the most frequently chosen primary area (39% and 35% respectively) for review while Area D and E sharply declined to 7% and 12% respectively by the third 4R Activity. The shift in trends for the primary area of improvement suggests that students began to emphasize feedback to ensure peer's NGSK products had all three necessary parts (claim, evidence and reasoning) over English language writing norms. This kind of feedback shift represents a deeper understanding of what a high quality NGSK product looks like as compared to more superficial emphasis on English Language grammar.

b. Characteristics of Reviews Provided by the Students

Nine salient characteristics emerged from the student reviews across all the classrooms and across the three review activities. In Table 3.4, the characteristics along with a representative example for the characteristics are provided. The representative examples are presented as direct quotes from students' reviews. In cases where clarification to the example was needed, the clarification is provided in square brackets "[]" in italicized font below the example.

Table 3.4

Representative examples for the review characteristics

Characteristics	Example Direct Quotes from Student Reviews
1. Suggestion includes only praise	“You did good”
2. Suggestion includes supportive language	“You got the right idea but you're not putting it in the best way so I suggest you add…”
3. Inconsistency between area of improvement & suggestion given	“Area C: Explanation/Prediction does not have specific reasoning Suggestion: In your reasoning you need to include more on how much the temperature rose in the graph.” *[The suggestion pertains to evidence as evidence in the curriculum was defined as data drawn from graphs, maps, tables, charts etc.]*
	“Area D: Explanation/Prediction needs stronger scientific words Suggestion: You need to check spelling of Montana and the animals you mentioned.” *[Suggestion is about spelling errors]*
4. Explicitly identifying where to make the improvement within the area selected	
a. Paraphrasing	“In your evidence you stated that Montana is too cold, you did not state how cold it will be.” *[The reviewer paraphrased a part of the author’s evidence. The direct quote from the author’s NGSK product pertaining to the paraphrased text was. “Most of the time Montana’s temperature is very cold ”]*
b. Direct Quotes	“When you said "on the graph it shows the line going up," I don't know what line you are talking about, or even what graph you are talking about.” *[The reviewer directly quoted a sentence from the author’s NGSK product (“on the graph…”) to draw attention to that particular sentence.]*
5. Specific Suggestion Given on how to make improvements	
a. Example	“In the reasoning, you need to talk about the predator-prey relationship for fox and the shrew.” *[The predator-prey relationship is given as an example]*
b. Hints	“You could use another source to really make the evidence solid, (hint: the simulations we did in lesson 5 that shows what happens to sun's energy when GHGs increase. That can be a good source of additional evidence)…”

	[Use of the word "hint". Only reviews with the explicit use of this word were coded as hints]
c. Sentence replacements	"In your evidence, include where you saw that the brown squirrel can live in Canada. Explain what you specifically saw that made you conclude on what you did. You could add something like: "In the future, the brown squirrel will live in Canada because on the prediction map Canada was the highlighted region." Somewhere along those lines."
6. Providing the reason for the suggested improvement	"In your reasoning and evidence you referred to the species distribution map as a "graph/map". Replace the words "graph/ map" with temperature map. I am suggesting this because this specific map shows the location of the animals which is called a species distribution map."
7. Future advise	"Pro Tip: I strongly suggest that for next time never put ";)"in anything you are typing that is supposed to be sophisticated." *[Advice given on not using emoticons in school or scientific writing (e.g., ";)")]*
8. Suggestion on Labeling/ Prefacing parts of the NGSK product.	"Make sure you start your evidence with My evidence is."
9. Direct use of the suggestion template provided in the suggestions column of the Review Sheet	"Please look at information provided (example: map) to come up with more specific evidence. In your evidence, include the following details: …" *[The suggestion template corresponding to the above quote was "Please look at information provided (example: maps, graphs) to come up with more specific evidence. In your evidence, include the following details: _________"]*

c. Observed Trends in Peer Reviews over Three Review Opportunities

The trends in the characteristics of student reviews were determined by calculating the frequency (in percentages) of the different characteristic observed in a review. The trends were determined for the characteristics discussed in Table 3.4. The trends of students with IEPs and students classified as ELLs are not discussed separately as there were no observed consistent trends within the ELL and IEP populations to need summarization separately.

Table 3.5 presents the frequencies of occurrence of the review characteristics for the three review opportunities that were provided as part of the 4R Activity. Both the primary and additional suggestions given during peer review were considered in the count of the total number of reviews. Depending upon the number of additional reviews given, this number varies across the three review opportunities within a class. Additionally, the total number of reviews per classroom is provided for each review opportunity.

The frequency of occurrence of a particular characteristic in a classroom is presented as a percentage of the total number of reviews considered for that particular round of reviewing. For example, in Ms. Lata's class the characteristic of 'using supportive language' had a percentage of 13% in the first review opportunity. This percentage is based on the 31 reviews that comprised the data for the first review opportunity. Thus in Ms. Lata's class 13% of the reviews had the use of supportive language in the first review opportunity.

Given that a review could possess multiple characteristics, the percentages do not add up to a 100%. Instead they provide a sense of the presence of that particular characteristic in the data for that class. For characteristics that can be further divided into sub-characteristics (e.g., “Specific Suggestion Given on how to make improvements” has three sub-characteristics of examples, hints and sentence replacements) the percentage corresponding to the sub-characteristics add up to a 100% and reflect the presence of the sub-characteristic in the characteristic. Additionally, a particular review possessed only of the three sub-characteristics and therefore the sub-characteristic adds to a 100%. For example, in Ms. Lata’s class for the first review opportunity, the characteristic of ‘Specific Suggestion given on how to make improvements’ in student reviews was observed in 42% of the reviews. Within this characteristic providing ‘examples’ in the suggestions given by students constituted 90%, hints constituted 0% while sentence replacement made up 10% of the data on the characteristic “Specific Suggestion Given on how to make improvements”. The percentages of the sub-categories are italicized in Table 3.5.

Table 3.5

Review Characteristics and their percentage of occurrences for each classroom for the 4R Activities

Class	**Ms. Lata (N=17)**			**Ms. Asha (N=17)**			**Ms. Priya (N=33)**			**Ms. Laila (N=30)**			**Mr. Amar (N=76)**		
4R Activity	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**
Total reviews in each 4R Activity (includes primary &additional reviews)	**31**	**32**	**34**	**30**	**33**	**34**	**53**	**62**	**63**	**50**	**52**	**57**	**131**	**133**	**133**
Characteristics	All cell values below are in percentages (%)														
1. Inconsistency b/w area of improvement &suggestion	19	13	9	23	15	6	15	5	2	30	19	7	2	0	0
2. Only praise	10	6	6	17	9	3	28	8	5	26	15	11	11	8	4
3. Using Supportive language	13	22	35	17	30	38	4	13	21	6	33	32	6	18	22
4. Using Suggestion Template	39	19	9	37	24	15	34	26	16	34	29	26	14	13	4
5. Providing the Reason for the suggested improvement	6	19	24	3	12	25	0	11	24	12	23	25	2	5	19
6. Future advice	13	6	15	7	15	9	0	11	14	11	23	23	2	6	10
7. Explicitly identifying where to make the improvement within the area selected (Total)	42	78	76	40	64	79	30	56	68	24	38	57	27	50	74
Paraphrasing*	*77*	*40*	*27*	*83*	*57*	*33*	*88*	*34*	*28*	*83*	*50*	*26*	*78*	*36*	*39*
Direct Quotes*	*23*	*60*	*73*	*17*	*43*	*67*	*12*	*66*	*72*	*17*	*50*	*74*	*22*	*64*	*61*
8. Specific Suggestion Given on how to make improvements (Total)	65	84	91	60	76	94	25	53	75	24	39	58	23	68	84
Examples*	*90*	*44*	*38*	*83*	*52*	*38*	*92*	*36*	*22*	*75*	*45*	*22*	*73*	*33*	*22*
Hints*	*0*	*37*	*52*	*11*	*40*	*50*	*0*	*46*	*55*	*25*	*40*	*45*	*10*	*41*	*50*
Sentence Replacements*	*10*	*19*	*10*	*6*	*8*	*12*	*8*	*18*	*23*	*0*	*15*	*35*	*17*	*26*	*29*
9. Suggestion on Labeling/ Prefacing sentences — Claim	2	9	6	1	12	15	0	0	0	0	0	0	0	0	0
Reasoning	2	22	13	0	36	18	0	29	11	0	23	11	0	21	9
Evidence	1	18	13	0	30	15	0	22	6	0	15	10	0	19	3

* Percentages for sub-categories add up to a 100% as they present the percentage breakdown of that particular category into the sub-categories.

It was observed that with each review cycle the inconsistency between the area of improvement and the suggestion decreased for all classes. In this category, the majority of the inconsistencies pertained to students selecting Area B (evidence) or C (reasoning) as an area of improvement and providing suggestions that did not correspond to that particular area. Additionally this also encapsulated Area D (suggesting scientific terminology) where students selected this area and provided suggestions pertaining to spelling errors in the NGSK product. Excluding Mr. Amar's class, where the inconsistency between the area of improvement and the suggestion was 2% in the first review, for all the other classes the range for inconsistency between area of improvement and the suggestion was between 15% -30%. This percentage dropped to between 6-9% for all the classes.

The characteristic "only praise" in the reviews saw a decline (became as low as 3-6% across all classes) by the third review activity while the use of supportive language when providing suggestion increased with each successive review activity and became as high as 38% (Ms. Asha's Class). For Ms. Priya and Mr. Amar's class, the use of supportive language rose to around 20% by the last review activity. The use of the suggestion template also declined with each review activity, except for Ms. Laila's class where it only slightly declined from 36% to 26%. One reason to explain the consistent use of the suggestion template in Ms. Laila's classroom is that she heavily emphasized its use. In discussions with Ms. Laila she revealed that in her view the use of the suggestion template allowed for the maintaining of quality in each review, as the suggestion template pre-defined certain parameters that the review had to contain. Additionally Ms. Laila expressed that the use of the template helped her students with IEPs to more easily provide review and in an effort to maintain the anonymity of the reviewers during the reviewing process, she encouraged all students to use the template.

The characteristic of “providing the reason for the suggested improvement” increased over time. This increase was most apparent for Ms. Priya’s classroom where the first review activity had 0% of the reviews including a reason for the suggested improvement and had 24% of the students providing a reason for the suggested improvement by the last review activity. During the first review process, it was discovered that many students’ reviews failed to clearly articulate why the proposed changes were suggested. Discussions with the teachers also revealed that students expressed not making changes in the revision process because they were unclear why the changes were required. Given this information, the scientist’s review attempted to explicitly articulate the reason for any suggested change, wherever possible. Additionally, classroom observations during the 4R Activity also showed that the teachers (especially Ms. Priya and Ms. Asha) reminded students to articulate the reasons for the suggested improvement. These could be potential reasons for the sharp increase in this area with each subsequent review activity.

As for the characteristic of reviewers providing advice for future writing of NGSK products, it was observed that this characteristic remained somewhat consistent across the review activities for the classes. However, for Ms. Laila’s class it almost doubled from the first review activity (11%) to the second (23%) and then remained constant. In Ms. Laila’s class there was heavy emphasis on the use of Standard English and following norms of professional writing, such as avoiding the use of emoticons and providing advice on conducting a spelling check before submitting the response.

Classroom observations revealed that Ms. Laila stressed the use of Standard English and conducting a spelling check before submitting the NGSK product. Advice similar to what Ms. Laila stressed in the classroom was seen in student reviews.

When examining the trends pertaining to the characteristic "Explicitly identifying where to make the improvement within the area selected", it was observed that over time students explicitly identified where the improvement had to be made. The reviews from Ms. Lata and Ms. Asha's classrooms had over 40% of the students explicitly identifying where the improvement had to be made in the first review activity. Their percentage went up to around 77% by the last review activity. The other classes began somewhere in between 24-30% reviews that explicitly identified the location of where the improvement had to be made and went up to 57-74% by the last review activity. By the last review activity all classes had similar percentages for this category. For example, In Mr. Amar's class 27% of the reviews explicitly identified the location of where the improvement had to be made in the first review activity and by the last review activity the percentage of 74% was comparable to Ms. Lata and Ms. Asha's class that had begun at 40% reviews identifying where the suggestion had to be incorporated.

The explicit identification of where the suggestion had to be incorporated was either done through paraphrasing or by direct quotations from the NGSK product that was being reviewed by the reviewer. Paraphrasing was much more prevalent than the use of direct quotations in the first review activity but by the last review activity this trend reversed with direct quotations had become much more prevalent than paraphrasing and while the percentage of direct quotation went up considerably.

Providing specific suggestions on how to make improvements also increased with each review cycle. The suggestions took the form of students *providing examples* for a concept (e.g., a predator-prey relationship), *hints* and *sentence replacement* that could be used. In the student's reviews, it was seen that both hints and sentence replacement were the least widely used form of providing suggestions on improvements, with most suggestions including examples only. The frequency of hints grew considerably in the second and third review activity. Sentence replacements increased for some classes over the three review activities (especially in Ms. Laila and Ms. Priya's classroom) while for Ms. Lata and Ms. Asha's class they remained low. Ms. Lata and Ms. Asha both had used hints in their classrooms when teaching the curriculum to the students. Thus the students were more used to using hints than sentence replacements. With respect to the use of examples, there was a decline in its use all classes except that of Ms. Lata and Ms. Asha where their decrease was less prominent as compared to the other classes. One explanation for this trend is that Ms. Lata and Ms. Asha's classes used the adapted version of the 4R Activity in their ELA class, they used hints and examples in their review. This might explain why the use of hints and examples was so high for these classes.

When it came to prefacing parts of the NGSK product, it was observed that over time students suggested prefacing parts of the NGSK product with either a label (e.g. "Reasoning:") or a sentence (e.g., "my reasoning is…"). It is to be noted that in the first review cycle, the NGSK product that was being reviewed was written in a pre-partitioned template with designated space for the three parts of the NGSK product. In the curriculum, these scaffolds were removed halfway through the curriculum and students were encouraged to write in paragraph format. That could account for extremely low percentage for this category in the first review cycle. The percentage of prefacing in reviews sharply increased in the second review activity and then decreased again in the last review activity. The decrease in the last review activity was more apparent for Ms. Priya, Ms. Laila, and Mr. Amar's classes. Perhaps this was because the pre-partitioned template to write the NGSK product had only been recently removed in the curriculum and students were adjusting to both writing and reviewing with NGSK products written in paragraph format. Additionally, classroom observations revealed that teachers encouraged students to continue using labels or prefacing parts of the NGSK product to indicate what it referred to. Some teachers (Ms. Laila, Ms. Lata, Ms. Priya) expressed that by doing so students would remember to include all parts.

In Table 3.6 it is seen that prefacing the claim was only observed in Ms. Lata and Ms. Asha's classrooms. This is because they emphasized that all parts of the NGSK product must be prefaced. Additionally they also requested that the suggestion on prefacing also be included in the scientist's review. These might be some reasons to explain the presence of prefacing parts of the NGSK product in students' own reviews and the continued presence of the prefacing suggestion in the last review activity. It is unclear, however, if this was solely due to the presence of a similar suggestion in the scientist's review or if it was due to the emphasis placed by teachers during their instruction, or both.

Another reason for the overall decline in prefacing suggestions across the classrooms could be due to the inclusion of prefacing suggestions by the authors of the NGSK product. In most of the NGSK products in the last 4R Activity, many authors had already prefaced the parts of the NGSK product and thus the reviewers did not need to include this suggestion in their review.

In an effort to determine some factors that might have contributed to an increase in the frequency of characteristics, five scientist's reviews from each class were coded. This was done to see if there was an uptake from the scientist's reviews for some of the characteristics that were emerging in the student reviews. As outlined in the earlier sections, the scientist's reviews were customized for each class and attempted to use the vocabulary that the teachers used to explain concepts in the class (e.g., "Reasoning is the *why* of the explanation"). The intention behind using class specific terminology in the reviews was to increase the uptake of the scientist's review by the students during revision.

The scientist's suggestions also included aspects that teachers wanted their students to focus on. For example, Ms. Lata and Ms. Asha emphasized that the scientist's reviews should mention that students preface parts (claim, reasoning, evidence) of the NGSK product with "my claim is; my reasoning is; my evidence is". Thus for these classes, the scientist's reviews included suggestions to students about prefacing parts of the NGSK products with appropriate labels or sentence starters to signify what part (claim, reasoning, evidence) was being addressed in the sentence. Table 3.6 showcases the frequency of occurrence of the characteristics from student reviews that were found in the scientist's reviews.

Table 3.6

Characteristics and their frequency of occurrence in the scientist's reviews

Characteristics Observed in the Scientist's Review		Ms. Lata (N=5)	Ms. Asha (N=5)	Ms. Priya (N=5)	Ms. Laila (N=5)	Mr. Amar (N=5)
Explicitly identifying where to make the improvement within the area selected (Total)		5	5	5	5	5
a. Paraphrasing		2	1	1	1	2
b. Direct Quotes		3	4	4	4	3
Specific Suggestion Given (Total)		5	5	5	5	5
a. Hints		4	4	3	3	4
b. Sentence Replacements		1	1	2	2	1
Using Supportive language		5	5	5	5	5
Providing the Reason for the suggested improvement		3	4	4	4	3
Future advise		1	1	2	2	1
Suggestion on Labeling/ Prefacing parts of the NGSK product e.g., "my reasoning is…";	Claim	4	4	0	0	0
	Reasoning	2	3	0	0	0
	Evidence	3	3	0	0	0

In Table 3.6, the readers can see that in all instances the scientist's reviews explicitly identified the location where the edits needed to be made. For this direct quotations were used more frequently in the scientist's reviews than paraphrasing. With respect to specific suggestions both hints and sentence replacements were used with hints being more prevalent in the scientist's review. Hints were more frequently used in the scientist's review to mirror the supports offered in the curriculum. In the curriculum hints were presented as scaffolds for construction of NGSK products.

Characteristics in students' reviews that increased over time (see Table 3.5) suggest that there is an uptake from the scientist's review in the student reviews for many characteristics. For example, the increase of the use of hints over time (from approximately 9% to 50%), the use of supportive language (from approximately 10% to 30%), use of direct quotes as an identification of the location where the edits had to be made (from approximately 18% to 70%). Direct quotations were more prevalent in the scientist's reviews than paraphrasing (see Table 3.6) when directing the author to the area where the improvement had to be made. This could be one reason for an increase in direct quotes in students' reviews. Additionally, classroom observations during the 4R Activity did not show teachers emphasizing one form over another.

3. Reflection on Providing Peer Reviews

The reflection activity was designed for students to reflect on their peer reviewing skills by re-reading the review they provided to their peer and also reading the review provided by the scientist to their peer. The patterns observed in students' reflection indicated that students became increasingly reflective on how they could improve their peer reviewing skills.

The reflections undertaken by the students could be placed on a continuum where moving from left to right lead to a progressive increase in the amount of reflection that was undertaken by the student. To show the progressive increase in reflection, each category of reflection was assigned a number that increased as one moved from left to right. This created a total of seven steps on the reflection continuum. A diagram showing the reflection continuum is shown in Figure 3.3.

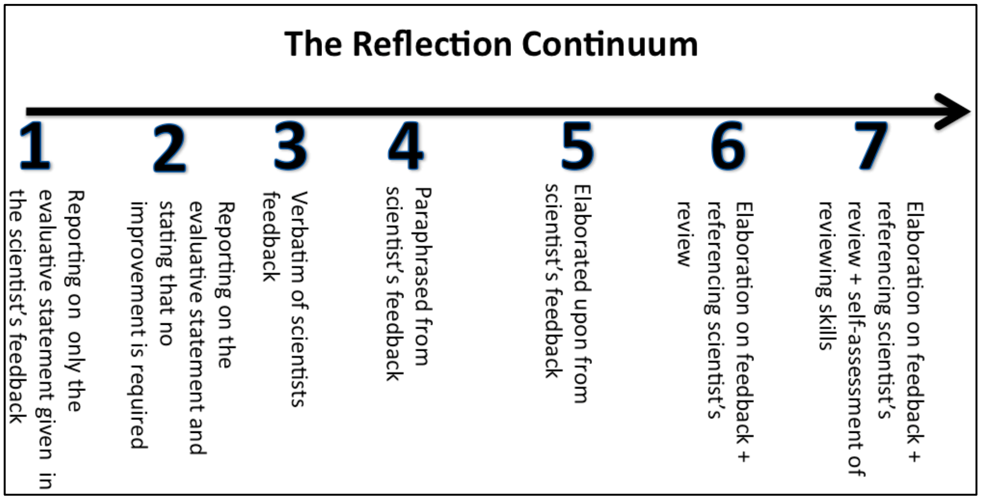


Figure 3.3. The Reflection Continuum to categorize students' reflections.

Beginning from the left hand side of the continuum, in the first step it was seen students only wrote about the evaluative statement given by the scientist on their review performance (e.g., "I did a good job"). In the second reflective step, in addition to writing an evaluative statement the student mentioned that no further improvement in peer reviewing skills was required because of the evaluation that had been received on the reviewing skills. Thus the first two types of reflections did not include the aspect of improvement that could be made in the reviewing skills.

Moving along the continuum, students began to reflect on how to improve their peer reviewing skills. The reflections undertaken by students who were only based on the scientist's feedback on his/her reviewing skills ranged from reflections that stated the scientist's feedback in verbatim (step 3) to an elaborated reflection on the reviewing skills based on the scientist's feedback to the reviewer (step 5). In the elaborated reflection, the student unpacked the scientist's feedback by providing supporting examples from their own review.

The sixth step on the reflection continuum included reflections that referenced portions of the scientist's review in addition to reflecting on the feedback given by the scientist to the reviewer. The last step on the reflection continuum were reflections that included a self-assessment by the student on his or her reviewing skills in addition to reflecting on the scientist's feedback and referencing the scientist's review. In this type of reflection the learner first self-assessed their own reviewing skills and then reflected on how to improve their reviewing skills by incorporating the scientist's feedback and provided supporting examples from the scientist's review. Refer to Appendix D for the reflection categories along with examples from student reflections.

a. Trends in Students' Reflections.

In this section, the trends across the participating students reflections are discussed. The trends are presented for each teacher's class to facilitate the comparison of trends across classes and allow for the discussion of possible reasons for observing a trend for a particular class. The trends of students classified as ELLs is also discussed separately wherever a pattern was seen to emerge within or across classrooms. The trends for students with IEPs were not summarized separately due to a lack of consistent trends within a class or across the classrooms.

When examining the overall trends, it was seen that for the first reflection activity, the most frequent type of reflection could be placed on the left hand side of the reflection continuum. The frequency of these reflections slightly declined by the third reflection activity. In the last reflection activity, the most frequent type of reflection could be placed on the right hand side of the reflection continuum, in steps 6 and 7. Thus students' reflections went on to become progressively reflective by the third reflection activity. The overall trends indicated that students’ reflections became more elaborate, encompassing different characteristics of reflection by the last reflection activity. The observed trends in student reflection are presented in Figure 3.4 and elaborated below.

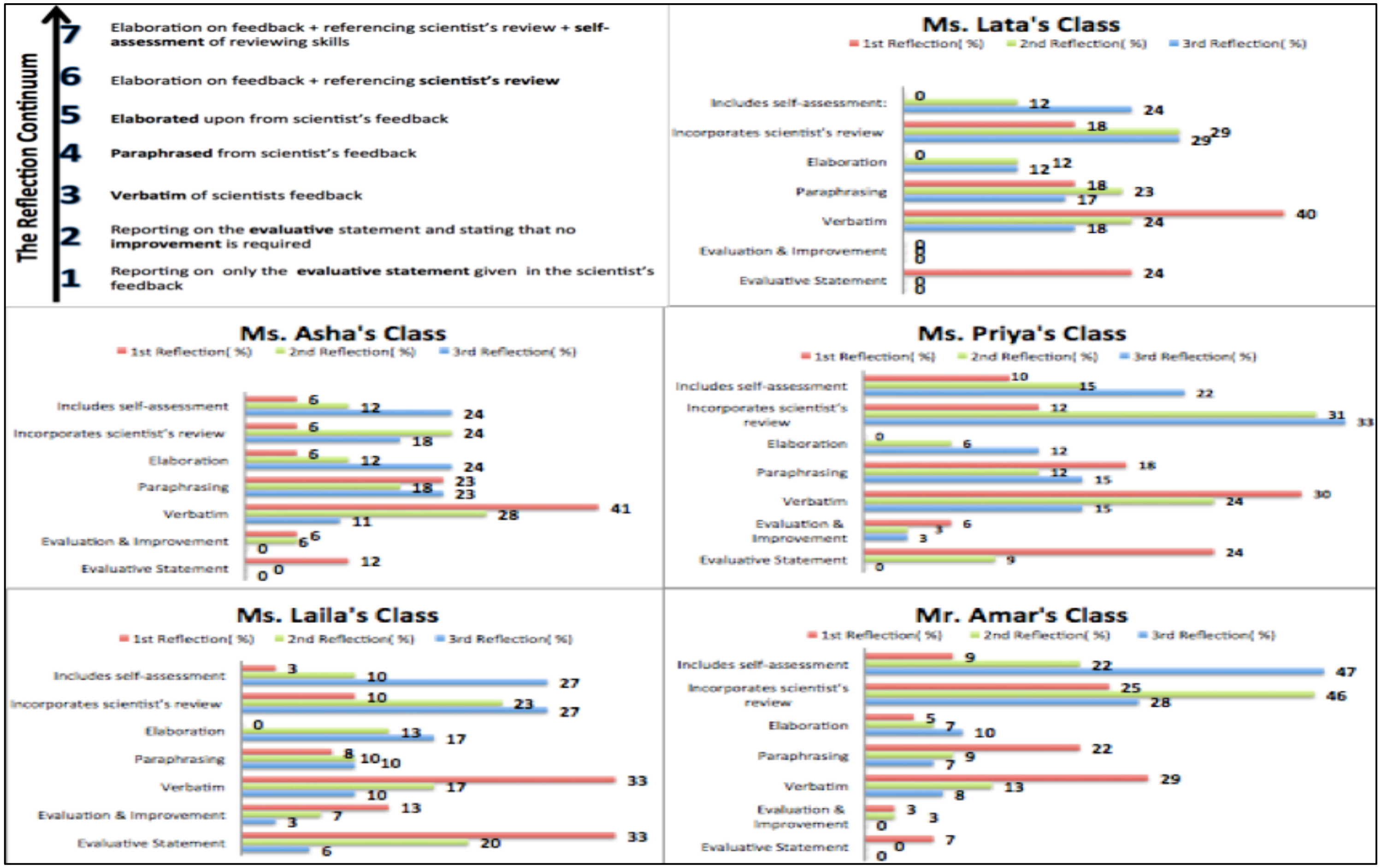


Figure 3.4. The trends observed in students' reflections.

As shown in Figure 3.4, there was a predominance of two types of reflections in the first reflection activity. The first type of reflection consisted of reporting on the evaluative statement given by the scientist to the reviewer (step 1on the reflection continuum) and the second was stating the scientist's feedback verbatim in the reflection (step 3). Reflections that comprised only evaluative statements sharply declined beginning from the second reflection activity (e.g., decline in evaluative statements from 24% in the first reflection activity to 0% in the second and third reflection activity in Ms. Lata's class). In the second reflection activity, students reporting on the scientist's feedback on the reviewing skills, which was taken verbatim from the scientist's feedback (step 3) continued to be a dominant trend (one of the highest percentage value) in the reflection across the schools (e.g., see Ms. Lata (24%), Ms. Asha (28%) and Ms. Priya's classes (24%)) However, the most dominant trend observed was reflections that elaborated on the scientist's feedback and referenced the scientist's review to the peer (step 6 on the reflection continuum). For example, in Ms. Lata's class 29% of the reflections were categorized as step 6 reflections. Similarly in Ms. Asha's class 24% of the reflections were categorized as step 6, while 46% of reflections were step 6 in Mr. Amar's class. The other types of reflections seen in the second reflection activity feel between steps 3 and 6 on the reflection continuum.

In the last reflection activity over 50% of the reflections were categorized as step 6 or step 7 on the reflection continuum. Across the three reflection activities, there was a decline in reflections categorized as steps 1 through 4 on the reflection continuum by approximately three times (e.g., For Ms. Laila's class the decline was from 21% to 7%).

Thus reflections began to get progressively more reflective over the course of three reflection activities.

As a reminder, students classified as ELL were present in only Ms. Lata, Ms. Asha and Mr. Amar's classes. They showed trends that were consistent with the students from their class in the three 4R Activities. In the first reflection activity, all of their reflections were categorized as step 3 (verbatim). In the second reflection activity, reflections categorized under step 3 continued to be predominant but there was a small presence of reflections that were categorized under step 4 (paraphrased) and even step 5 (elaborated). These reflections constituted 5% of the reflections done by ELLs across the three classes. In the last reflection activity, step 3 reflections continued to exist but their percentage dropped from 95% in the second reflection activity to 40% in the last reflection activity. Reflections categorized under step 4 (paraphrased) also rose to 15% while the reflections categorized under step 5 (elaboration) rose to 30% across the ELL students in the three classes. The remaining 15% of the reflections of ELLs were seen categorized under step 6 and step 7.

b. Additional Aspects in Students' Reflections

There were four additional aspects observed in student reflections that were seen in the reviews that were present for steps 5, 6 and 7 of the reflection continuum. These aspects were (1) reflecting on specific improvements the student would make to his/her present review, (2) reflecting on the construction of the NGSK product, (3) linking their reviewing skills to the construction of the NGSK product, and (4) reflecting on the purpose of the 4R Activity. All these aspects were categorized as "additional aspects" because they included aspects of reflection that were not directly related to the reflection activity prompt, which asked students to reflect on their peer reviewing skills (See Appendix E for examples on additional aspects from student reflections). The observed trends in the additional aspects are presented in Figure 3.5.

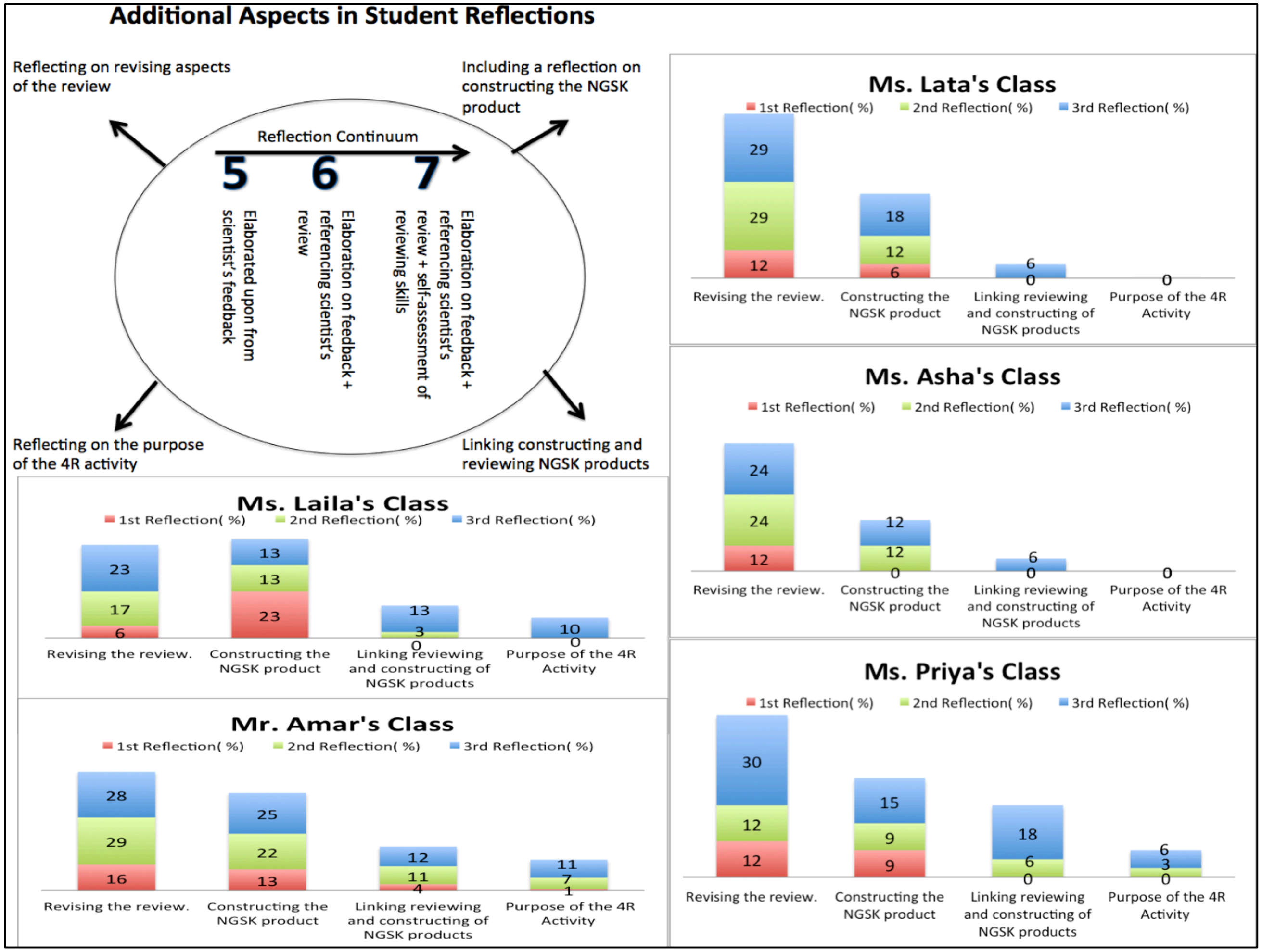


Figure 3.5. Trends pertaining to the "additional aspects" observed in students' reflection activities.

For the trends in the four additional aspects (Figure 3.5), it was observed that a reflection including aspects on revising the review increased with each reflection activity. Additionally, reflecting on constructing the NGSK product was also observed, with it becoming more frequent in the second and third reflection activity. Amongst the additional aspects, 20% of reflections of ELL students had these two aspects present in the reflections. Out of these 15% included reflections on constructing the NGSK product. Another aspect that was observed was the linking of the construction and reviewing of NGSK products. This became more prevalent in the last reflection activity. The aspect of reflecting on the purpose of the activity was not observed for all classrooms. This category was most frequent in the last reflection activity. Perhaps the students treated the last reflection activity as an opportunity to communicate their thoughts on the entire 4R Activity.

Discussion

This study proposed an approach to support middle school students in the construction NGSK by engaging them in the activity of receiving and providing critique on their NGSK products. Five salient themes emerged from the findings that are discussed in this section

1. Students reviews evolved from focusing primarily on writing to salient components of NGSK

The process of revision was based on the feedback students received. Research studies have shown that elaborated and specific feedback is more helpful to the receiver of feedback than concise and general feedback (Hattie & Timberley, 2007; Shute, 2008). Peer review characteristics over the three review activities indicate that reviews became more elaborated and specific. Thus in this study students' reviews might have become more helpful to the author of the NGSK product and this might have also contributed to the improvements in the revised NGSK products. In the peer review activities, students first began with reviewing only the writing aspects of the NGSK product, which was in line with research on student writing (e.g., Flower, Hayes, Carey, Schriver & Stratman, 1986). However over time students began to also provide review on the content of the NGSK product. However feedback pertaining to the writing aspects was still prevalent in peer reviews but it began to occur as an additional area of improvement. From the second review activity onwards, students were seen providing review on an additional area of improvement, which was not done in the first review activity. For additional improvements, they began to provide suggestions on the inclusion of scientific terminology to make the NGSK product strong.

An emphasis on the writing aspects in earlier reviews could have been due to the documented challenges faced by students in constructing evidence-based explanations. For example, students struggles with what counts as evidence and reasoning in evidence-based explanations (Sadler, 2004; Sandoval, 2003; Bell & Linn, 2000) could have contributed to them not choosing those areas in the beginning. As they progressed through the curriculum they engaged deeply in both the core concepts as well as the practice of constructing NGSK products. A progress in learning pertaining to these two areas, they might have then contributed to students providing reviews related to the specifics of constructing NGSK products.

A shift in focus from writing aspects in the review could have also occurred due to factors such as, students learning the importance of reviewing other aspects of the

NGSK product, improved writing quality of constructed NGSK products over time, and providing review on aspects that students themselves found helpful during revision. As students progressed along the curriculum and constructed NGSK products the criteria of what constitutes a strong NGSK product might have become clear to the students and with that their peer reviews also began to include areas of improvement besides writing aspects. Additionally successive peer review activities might have also contributed to students understanding the criteria of a strong NGSK product. As outlined above, the writing quality of the NGSK products also improved over time as students became more adept in constructing NGSK products and also more aware that they were writing for an audience besides their teacher. Classroom observations revealed that teachers reminded students to conduct spelling-checks and re-read their NGSK product before submission, which might have also contributed to reviewers becoming less focused on writing aspects. However feedback on scientific terminology, which increased over time in the reviews can also be considered a part of the writing process as students suggested words and phrases to be included by the author. This shows that students continued to emphasize the writing aspects of the NGSK product but this emphasis was not the prime focus of the reviews.

Contrasting this result with students classified as ELLs, it was observed that Area E (Other Improvement) continued to be the primary area of improvement. Additionally a large majority of ELL students chose Area D (scientific terminology) as an additional area of improvement. While non-ELL students' shifted their primary focus from writing aspects to more content-focused aspects when reviewing NGSK products, students classified as ELLs continued to place prime emphasis on the use of language and writing aspects of the NGSK products. The trends suggest that students might be providing review on aspects they would like to receive feedback on. Perhaps students first look for those areas of improvement that they find challenging in the construction of NGSK products. The trends on what areas of improvement students chose to review also provides insight into what students consider important when providing feedback. These observed trends reflect the importance of vocabulary acquisition for ELLs and also reflect the challenges faced by ELLs with academic vocabulary acquisition (see Marzano, 2004; Graves, 2006). Vocabulary knowledge is documented to be a critical factor in determining the school success of ELLs (Carlo, August, & Snow, 2005). Within vocabulary acquisition, academic vocabulary has shown to be the greatest challenge for ELLs (Graves, 2006). Academic vocabulary poses a challenge because it consists of words with precise meaning, which often differ from other meanings of the same words in a different context (Marzano, 2004). Additionally, academic vocabulary has shown to be central in comprehending informational texts in science and mathematics (Marzano, 2004; Graves, 2006). Given the varied meanings of academic vocabulary and the dependence of academic vocabulary in comprehending texts in science, ELLs might be placing more emphasis on acquisition of academic vocabulary.

2. Students' reviews became more detailed and specific over time

A salient result in the characteristics of student reviews was that over time there was an increase in the elaboration (from approximately 30% to 75%) and specificity (from approximately 40% to 80%) in the reviews provided by students. Students began to provide specific suggestions, identified where improvements had to be made and also provided the rationale for the suggested improvement. Multiple factors might have contributed to the increase in elaborated and specific peer reviews. For example, classroom observations revealed that teachers emphasized the need for specificity in the peer reviews, which might have served as a reminder for students to provide specific peer reviews. Additionally reading the reviews students received in previous 4R Activities might have also contributed. Another important factor that might have helped in reviews becoming more elaborate was the reflection on peer review skills. Given the high uptake from the scientist's review in student reflections, the reflection activity could have made expert strategies of reviewing available to the learner. This is in line with research that shows that having access to an expert's strategies promotes learner's strategies to be more in line with the expert (Pederson & Liu, 2002).

3. Students were able to incorporate feedback during the revision process that lead to an increase in their scores

The purpose of the revision activity was to help students improve their NGSK product by incorporating the feedback that they received from their peer and a scientist. In the trends observed over the three revision activities, four salient results emerge. They are as follows:

(1) Increase in the percentage of students who demonstrated a score level improvement and decrease in the percentage of students who did not demonstrate a score level improvement on their revised NGSK product.

(2) Increase in the percentage of students who began at lower score levels (e.g., level 1 and 2) and achieved proficiency levels on their revised NGSK product.

(3) A large proportion of the ELL population achieved proficiency levels on their revised NGSK products

(4) Students with IEPs were seen achieving proficiency levels on the revised NGSK products only in classes where support was provided by the teacher and were less successful in attaining proficiency levels in classes where they worked independently with intermittent support from the teacher.

The observed increase in the score level improvement on the revised NGSK product suggests that students became more receptive to the feedback offered to them and began to incorporate the feedback in their revised NGSK product. This could be seen by the decrease in the percentage of students not making any changes to their original NGSK product. Besides, becoming less resistant to the process of revision with each successive 4R Activity, students might have also become more effective in incorporating feedback over time. Ms. Lata and Ms. Asha’s classes provide an insight into this reasoning. These two classes had the highest percentage of students achieving proficiency levels on their NGSK product in the third 4R Activity (78% on Ms. Lata’s class & 60% on Ms. Asha’s class). In the two classes, peer review and revision of NGSK products was implemented seven times outside the 4R Activity. These students therefore had more opportunities to learn how to incorporate the feedback in their revision process. Additionally, the students also used an adapted version of the 4R Activity in their ELA class four times, which provided them with additional practice in effective use of feedback. The frequent use of peer feedback could have also helped students in these two classes to see the credibility of the feedback and be less resistant to the process of revision. This is also supported by research (Boud, 2000;Orsmond, Merry & Reiling, 2000) that suggests that the creation of a classroom climate where providing and receiving of peer feedback is a regular part of the teaching and learning processes makes students more receptive to receiving feedback and also enables them to provide constructive feedback.

4. More students starting from lower score levels were able to reach proficiency levels on their revised NGSK products

The results on the attainment of proficiency levels on the revised NGSK product suggest that the 4R Activity helped in narrowing the achievement gap between students as learners beginning at lower score levels were also able to attain proficiency. The findings indicated that students who began at a score level of at least 3 were most likely to be seen achieving proficiency levels. However this trend changed over time and students with lower score levels on the original NGSK product were also able to attain proficiency on their revised NGSK product. Examples of this can be seen in Ms. Priya and Ms. Laila's class where students beginning at levels of 1 or 2 began to reach proficiency levels on their revised NGSK products. One rationale to explain this trend is that earlier on, students were learning how to incorporate the feedback that could take them to a proficiency level and this was easier for students who began at higher levels because they had to make lesser number of changes to achieve proficiency. On the other hand students who began at a level 1 or 2 had to incorporate a lot of changes in their revised draft for it to reach a proficiency level. With continued practice on incorporating feedback students who began at a lower score level (level 1 or 2) on their original NGSK product were also able to make substantive improvements in their score levels to achieve proficiency levels.

With respect to the proficiency levels of ELLs, a majority of the ELLs were able to reach proficiency levels. All ELLs were provided continued support by their teachers during the 4R Activity, which could have aided in them achieving proficiency levels. This support was seen pertinent to helping ELLs because of the language demands placed on students in the 4R Activity. Teachers guided the ELLs during the revision process by reading the feedback to them and helping with elaboration wherever it was needed. These accommodations afforded the ELLs with the necessary means to reach proficiency levels. This is in line with research that shows that students with limited language proficiency consistently underperform on tests and changes to the language used in the assessment has shown to improve the performance of students with limited language proficiency (Aiken, 1972; Kintsh & Greeno, 1985).

Students with IEPs did show improvement in score levels, but were not frequently seen reaching proficiency levels. The students with IEPs were primarily in Ms. Laila and Ms. Priya's class where students worked independently or were provided assistance by another teacher. Due to a lack of direct communication between the author of this paper and the assistant teacher, it is unclear what support was provided to the students with IEPs in Ms. Priya and Ms. Laila's class during the 4R Activity. The learning needs of students with IEPs can vary considerably and it is unclear what challenges the students with different IEPs faced during the revision activity. A small sample size and lack of information on the accommodations provided to the students with IEPs restricts the interpretation of the results. Therefore future work is needed to understand the challenges faced by students in different IEPs in the construction of NGSK products as well as the support provided during incorporation of the feedback during the revision process.

5. Student reflections became more elaborated over time

Results on student reflections reveal that over time the frequency of verbatim reflections sharply declined, while paraphrasing and elaborating on the scientist's feedback increased over time. Additionally, reflections that included self-assessment done by the student on his/her peer reviewing skills also increased over time. Students also began reflecting on improvements specific to the review that they gave along with reflecting on authoring NGSK products, while some students also linked their authoring and reviewing of NGSK products. This shows that by the third reflection, students had begun gaining some conscious access to their own learning strategies where their reflections began to include components of self-assessment of their own reviewing skills in light of the review of the scientist. Another example that shows students consciously reflecting on their reviewing strategies is when they began to link the authoring and reviewing process which might indicate that they were trying to articulate the reasons for why they were unable to use a particular strategy that they saw in the scientist's review. Some students were able to decontextualize from the review process during reflection when they reflected about the purpose of the 4R Activity. In these reflections they included a reflection on what they thought the review process added to their own learning.

The findings on both reflection and review suggest that when student reflections became more elaborated and went beyond re-iterating the scientist's feedback, this created opportunities for students to internalize the strategies of reviewing upon which they reflected. These findings reflect Vygotsky's (1978) ideas on how active collaboration in a community promotes internalization of strategies and knowledge by the participants of the community. Given this study did not look at the percentage uptake of student's own reflection in their next review activity due to the small number of review and reflection activities, this hypothesis cannot be verified. Future work that draws connections between reflection and its uptake in the subsequent reviews will provide evidence on the effectiveness of the reflection process in building the peer review skills of students.

The findings from this study provide evidence for how engaging in the 4R Activity not only helped students to attain proficiency levels but also developed their competency in the process of receiving and providing critique. The study paves way for future work in the design of classroom formative assessments that promote development of proficiency in NGSK through the use of activities that reflect the work of scientists and engineers.

References

Abedi, J. (2010). Research and recommendations for formative assessment with english language learners. In Andrade, H., & Cizek, G. J. (Eds.). (2010). *Handbook of formative assessment*. Routledge.

Aiken, L. R. (1972). Language factors in learning mathematics. *Review of Educational Research, 42*(3), 359–385.

Anderson, J. O., Lin, H. S., Treagust, D. F., Ross, S. P., & Yore, L. D. (2007). Using large-scale assessment datasets for research in science and mathematics education: Programme for International Student Assessment (PISA). *International Journal of Science and Mathematics Education*, *5*(4), 591-614.

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

Berland, L.K., & Hammer, D. (2012). Students' framings and their participation in scientific argumentation *Perspectives on scientific argumentation* (pp. 73-93): Springer US.

Boud, D. (2000) Sustainable assessment: rethinking assessment for the learning society. *Studies in Continuing Education, 22*(2), 151-167.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (Eds.). (1999). How people learn: Brain, mind, experience, and school. Washington, DC: National Academy Press.

Bridwell, L. S. (1980). Revising strategies in twelfth grade students' transactional writing. *Research in the Teaching of English, 14*, 197-222.

Brown, S., Rust, C., & Gibbs, G. (1994). *Strategies for diversifying assessment in higher education*. Oxford: Oxford Centre for Staff Development.

Budden, A. E., Tregenza, T., Aarssen, L. W., Koricheva, J., Leimu, R., & Lortie, C. J. (2008). Double-blind review favors increased representation of female authors. *Trends in ecology & evolution*, *23*(1), 4-6.

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of educational research*, *65*(3), 245-281.

Carlo, M.S., August, D., & Snow, C.E. (2005). Sustained vocabulary-learning strategies for English language learners. In E.H. Hiebert & M. Kamil (Eds.), *Teaching and learning vocabulary: Bringing research to practice* (pp. 137–153). Mahwah, NJ: Erlbaum.

Cohen, G. L., Steele, C. M., & Ross, L. D. (1999). The mentor's dilemma: Providing critical feedback across the racial divide. *Personality and Social Psychology Bulletin*, *25*(10), 1302-1318.

Chen, C., & Crawford, T. (1998, April). Scientists in the making: An ethnographic investigation of students' access to scientific knowledge in an elementary classroom. Paper presented at the annual meeting of the National Association for Research in Science Teaching, San Diego, CA.

Collins, A., & Brown, J. S. (1988). *The computer as a tool for learning through reflection* (pp. 1-18). Springer US.

Collins, A., Brown, J. S., & Newman, S. E. (1989). Cognitive apprenticeship: Teaching the crafts of reading, writing, and mathematics. In L. B. Resnick (Ed.), Knowing, learning, and instruction: Essays in honor of Robert Glaser (pp. 453–494). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

De Guerrero, M. C. M., & Villamil, O. S. (2000). Activating the ZPD: Mutual scaffolding in L2 peer revision. *The Modern Language Journal, 84*, 51–68.

Emerson, Fretz & Shaw (2011). *Writing ethnographic fieldnotes (2nd edition).* Chicago: University of Chicago Press.

Elliot, S.N., Kettler, R.J., Beddow, P.A., Kurz, A. (2010). Research and strategies for adapting formative assessments for students with special needs. In Andrade, H., & Cizek, G. J. (Eds.). (2010). *Handbook of formative assessment*. Routledge.

Erduran, S., & Jiménez-Aleixandre, M. P. (2008). *Argumentation in science education*. Florida State University-USA: Spinger.

Flower, L., Hayes, J. R., Carey, L., Schriver, K., & Stratman, J. (1986). Detection diagnosis and the strategies of revision. *College Composition and Communication, 37*(1), 16–55.

Gotwals, A. W., & Songer, N. B. (2009). Reasoning up and down a food chain: Using an assessment framework to investigate students' middle knowledge. *Science Education, 94*(2), 1-23.

Graves, M.F. (2006). *The vocabulary book.* New York: Teachers College Press.

Harding, A. D. (2010). How to Phrase Feedback in Peer Reviews for Nurse Authors?. *Advanced Emergency Nursing Journal*, *32*(4), 333-337.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*, 81-112.

Hambleton, R., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory*. London: Sage.

Jimenez-Aleizandre, M. P., Rodriguez, A. B., & Duschl, R. A. (2000). "Doing the lesson"or "doing science": Argument in high school genetics. *Science Education, 84*, 757-792.

Kelly, G. J., & Bazerman, C. (2003). How students argue scientific claims: A rhetorical-semantic analysis. *Applied Linguistics, 24*(1), 28–55.

Kelly, G. J., & Chen, C. (1999). The sound of music: Constructing science as sociocultural practices through oral and written discourse. *Journal of Research in Science Teaching*, *36*, 883–915.

Kelly, G.J., & Crawford, T. (1997). An ethnographic investigation of the discourse processes of school science. *Science Education, 81*, 533–559.

Kintsch, W., & Greeno, J. G. (1985). Understanding and solving word arithmetic problems. *Psychological Review, 92*(1), 109–129.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: Historical review, a meta-analysis and a preliminary feedback intervention theory. *Psychological Bulletin, 119*, 254–284.

Lee, H.S. (2003). *Scaffolding elementary students' authentic inquiry through a written science curriculum*. University of Michigan, Ann Arbor.

Linn, M.C. & Eylon, B.S. (2006). Science Education: Integrating Views of Learning and Instruction. In P. A. Alexander & P. H. Winne (Eds.), Handbook of Educational Psychology (2nd Ed., pp. 511-544). Mahwah, NJ: Lawrence Erlbaum Associates.

Lin, S. S. J., Liu, E. Z. F., & Yuan, S. M. (2001a). Web-based peer assessment: feedback for students with various thinking-styles. *Journal of Computer Assisted Learning*, *17*, 420-432.

Liu, N. F., & Carless, D. (2006). Peer feedback: the learning element of peer assessment. *Teaching in Higher education*, *11*(3), 279-290.

Lu, R., & Bol, L. (2007). A comparison of anonymous versus identifiable e-peer review on college student writing performance and the extent of critical feedback. *Journal of Interactive Online Learning*, *6*(2), 100-115.

Lu, J., & Law, N. (2012). Online peer assessment: effects of cognitive and affective feedback. *Instructional Science*, *40*(2), 257-275.

Lundstrom, K., & Baker, W. (2009). To give is better than to receive: The benefits of peer review to the reviewer's own writing. *Journal of Second Language Writing*, *18*(1), 30-43.

Marzano, R.J. (2004). *Building background knowledge for academic achievement: Research on what works in schools*. Alexandria, VA: Association for Supervision & Curriculum Development.

Maxwell (2013). *Qualitative research design: An interactive approach.* Thousand Oaks, CA: Sage.

McNeill, K. L., Krajcik, J. S. (2012). *Supporting grade 5-8 students in constructing explanations in science: the claim, evidence, and reasoning framework for talk and writing.* Boston: Pearson.

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

McConachie, S. M. (2010). Disciplinary Literacy: a principle-based framework. In McConachie, S. M., & Petrosky, A. R. (Eds.). (2010). *Content matters: a disciplinary literacy approach to improving student learning.* San Francisco: Jossey-Bass.

Meskill, C. (2010). Moment-by-moment formative assessment of second language development: ESOL professionals at work. In Andrade, H., & Cizek, G. J. (Eds.). (2010). *Handbook of formative assessment*. Routledge.

Moje, E. B. (2007). Developing socially just subject-matter instruction: A review of the literature on disciplinary literacy teaching. *Review of research in* *education*, *31*(1), 1-44.

National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas.* Committee on Conceptual Framework for the New K-12 Science Education Standards. Board on Science Education. Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

NGSS Lead States (2013). *Next Generation Science Standards: For States, By States.* Washington, DC: The National Academies Press.

Nicol, D., Thomson, A., & Breslin, C. (2014). Rethinking feedback practices in higher education: A peer review perspective. *Assessment and Evaluation in Higher Education, 39*, 102-122. doi: 10.1080/02602938.2013.795518.

Norris, S. P., & Phillips, L. M. (2003). The public understanding of science information: Communicating, interpreting, and applying the science of learning. *Education Canada*, *43*(2), 24–27, 43.

Orsmond, P. Merry, S. & Reiling, K. (2000) The use of students derived marking criteria in peer and self-assessment, *Assessment and Evaluation in Higher Education, 25* (1), 23-28.

Palincsar, A.S. (1998). Social constructivist perspectives on teaching and learning. *Annual Review of psychology*, *49* (1), 345–375.

Patton, M.Q. (2002). Qualitative Research and Evaluation Methods (3rd ed.). Thousand Oaks, CA: Sage Publications.

Pedersen, S., & Liu, M. (2002). The effects of modeling expert cognitive strategies during problem-based learning. *Journal of Educational Computing Research*, *26*(4), 353-380.

Piaget, J. (1970). Piaget's Theory. In L. Carmichel (Ed.), *Carmichael's Manual of child psychology.* Chicago: John Wiley and Sons.

Reiser, B. J. (2013, September). What professional development strategies are needed for successful implementation of the Next Generation Science Standards. In Paper written for the Invitational Research Symposium on Science Assessment, September (Vol. 24, p. 25).

Sandoval, W. A., & Millwood, K. A. (2005). The quality of students' use of evidence in written scientific explanations. *Cognition and Instruction, 23*, 23–55.

Shattell, M. M., Chinn, P., Thomas, S. P. and Cowling, W. R. (2010). Authors' and editors' perspectives on peer review quality in three scholarly nursing journals. *Journal of Nursing Scholarship, 42*,58–65.

Shanahan, T., & Shanahan, C. (2008). Teaching disciplinary literacy to adolescents: Rethinking content-area literacy. *Harvard Educational Review*, *78* (1), 40-59.

Shanahan, T., & Shanahan, C. (2012). What is disciplinary literacy and why does it matter?. *Topics in Language Disorders*, *32*(1), 7-18.

Shaw, V. N. (2002). Peer review as a motivating device in the training of writing skills for college students. *The Journal of College Reading and Learning*, *33*(1), 68-76.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*, 153-189.

Siero, F., & Van Oudenhoven, J. P. (1995). The effects of contingent feedback on perceived control and performance. *European Journal of Psychology of Education, 10,* 13–24.

Songer, N.B., Kelcey, B., & Gotwals, A. (2009). How and when does complex reasoning occur? empirically driven development of a learning progression focused on complex reasoning About Biodiversity. *Journal of Research in Science Teaching, 46*(6), 610 - 633.

Songer,N.B., & Gotwals, A.W. (2012). Guiding explanation construction by children at the entry points of learning progressions. *Journal of Research in Science Teaching, 49(2),* 141–165.

Songer, N. B., & Kali, Y. (2014). Science Education and the Learning Sciences as Coevolving Species. *Cambridge Handbook of the Learning Sciences.*

Songer, N.B., Dewey, T., Peters, V., Reicher, M., Kwok, A., Zaidi, S., Tupper, B., Hammond, G., Myers, P. (2014) *Climate Change and Impacts on Ecosystems: A Middle School Curricular Unit.* Ann Arbor, MI: The University of Michigan.

Songer, N.B., Zaidi, S., Newstadt, M. (2016). Classroom-based holistic assessment and evidence of middle school students' Next Generation Science Knowledge. Manuscript in preparation.

Stemler, S.E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating inter-rater reliability. *Practical Assessment, Research & Evaluation*, *9* (4).

Teo, A. K. (2006). Social-interactive writing for english language learners. *The CATESOL Journal, 18*, 160–178.

Tenopir, C., & King, D. W. (2004). *Communication patterns of engineers*. John Wiley & Sons.

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

Tseng, S. C., & Tsai, C. C. (2007). On-line peer assessment and the role of the peer feedback: A study of high school computer course. *Computers & Education*, *49*(4), 1161-1174.

Venables, A., Summit, R. (2003). Enhancing scientific essay writing using peer assessment. *Innovations in Education and Teaching International*, *40*(3), 281-290.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard Univ. Press.

Windschitl, M., Thompson, J., & Braaten, M. (2008). Beyond the scientific method: Model-based inquiry as a new paradigm of preference for school science investigations. *Science education*, *92*(5), 941-967.

Williams, S. M. (1993). Putting case based learning into context: examples from legal, business, and medical education, Journal of the Learning Sciences, *2*, 367-427.

Yore, L. D., Florence, M. K., Pearson, T. W., & Weaver, A. J. (2006). Written discourse in scientific communities: A conversation with two scientists about their views of science, use of language, role of writing in doing science, and compatibility between their epistemic views and language. *International Journal of Science Education, 28*, 109–141.

Chapter 4

Eliciting Middle School Students' Perspectives on Using a Classroom Formative Assessment that Simulates the Double-Blinded Review Process of Academic Journals

Introduction

The *Framework for K-12 Science Education* (NRC, 2012) along with the *Next Generation Science Standards* (NGSS) (NGSS Lead States, 2013) call for a shift in science education from an emphasis on learning and assessing primarily scientific facts to learning and assessing integrated knowledge where disciplinary content is learned through the use of the practices of professional scientists and engineers (NRC, 2012). As researchers begin to design learning environments that foster learning of integrated knowledge they need to ask the question, "What do we know about what students think when they engage in activities designed to foster learning of integrated knowledge?" This question becomes particularly important for classroom assessment interventions, given one role of classroom assessments is to elicit students' interests and experiences that may be relevant to fostering further learning (NRC, 2014). However, eliciting students' interests and experiences through classroom assessments tends not to be a research focus in K-12 science (NRC, 2014). Responding to this gap in research, this study presents students' accounts of their experiences related to participating in a classroom formative assessment called the *4R Activity* (see Chapter 1, this dissertation).

In this study, students' perspectives are documented pertaining to participation in the 4R Activity (see Chapter 1, this dissertation) that is embedded within an online middle school science curriculum on *Climate Change and its Impacts on Ecosystems* (Songer et.al, 2014). The curriculum is designed to support students in the construction of *Next Generation Science Knowledge* (NGSK) products that integrate the three dimensions of NGSS (NGSS Lead States, 2013) into one knowledge product (see Chapter 1, this dissertation). The following research questions guided the analysis reported on in this paper:

1. What are student perspectives on learning from participating in the 4R Activity?

2. What applicability do students see for the 4R Activity in science class, other classes and/or in their life outside school?

    a. How do they describe this applicability?

3. Do students draw connections between the 4R Activity and the activities undertaken by scientists, engineers and other professionals?

    a. What similarities or differences do they describe if any?

Relevant Background Information

This section highlights important research documenting the need for student perspectives in three areas: (a) learning from participation in an activity (b) the connections between in-school-learning and its applicability to their lives both in and outside school, and (c) the connections between the learning environment and the practices of scientists, engineers and other professionals.

1. Engaging in the Activity of Peer Review

Research has documented various affordances of engaging in the activity of peer review. For example studies have shown that the activity of receiving and providing feedback benefits the reviewer as well as the receiver of the feedback (see Lunstrom & Baker, 2009; Liu &Carless, 2012). Receiving feedback allows learners to undertake an active role in their own learning and provides them with an opportunity to monitor their learning (Butler & Winnie, 1995;Venables & Summit, 2003). Additionally, providing feedback exposes students to the different solutions adopted by their peers for a given problem, which in turn can allow students to consider the value of their contemporaries' work (Purchase, 2000). Additionally, providing feedback can also help students to identify both good and bad practices in their peers' work (Purchase, 2000). Furthermore engaging the peer review also mirrors the professional environment of receiving and providing feedback that students are likely to encounter in their professional workplace (Purchase, 2000).

However research also reports some challenges associated with using peer feedback activities in classrooms, specifically the problem of biased feedback arising out of stereotypes associated with gender, race or minority status. For example, Cohen, Ross and Steele (1999) discuss how students from underrepresented groups might feel that their feedback is based on a negative stereotype associated with the intellectual ability of their group rather than on their own intellectual merits. Biased feedback or the perception of bias in feedback can in turn hinder the uptake of feedback and result in some students withdrawing from doing the task. In this study the challenge of potential bias in feedback is addressed by introducing a double- blinded feedback system, where the identity of the author and the reviewer are hidden from each other. Given that the double-blinded review process was introduced to counter the potential problem of biased feedback it was important to document students perspectives on this aspect and if students thought that the double-blinded review facilitated unbiased feedback.

2. The Need to Understand Student Perspectives

Research supports the perspective that students might experience any given learning environment in a way that is different from how it was intended by the designer (see Elen & Lowyck, 1998; 1999). Therefore it is important to capture the perspectives of students to preserve the *local meanings* (Erickson, 1986) of an activity from the point of view of the participants. Additionally capturing student perspectives can provide some insight into the reasons for observed achievement gains, given students' perceptions of a learning environment can impact the effectiveness of the learning environment in promoting student learning (Entwistle, 1991). Furthermore, scholars (e.g. Bell, Bricker, Reeve, Zimmerman &Tzou, 2013) have also argued for the need of the science education community to better understand the ideas, perceptions and experiences of youth that engage with the learning environment that researchers design. They argue that understanding the perspectives of the youth involved is an important design consideration if designers want to create learning environments that are in sync with the learning needs of today's youth (Bell et al., 2013). This study is grounded in the argument that we need to gain insight into the ideas, perceptions and experience of learners to understand their thinking about aspects such as how they best learn, what types of activities are most engaging to them, and what connections they draw between the science learning in their classrooms, for example, and their lives. It is only by documenting learners' perspectives can we develop an understanding of their perceptions. Moreover, this research argues that the information gathered from learners' perspectives in turn could inform our design decisions on how to make learning more accessible and meaningful to the students.

a. Students' perspectives on learning. Scholars have argued that listening to students' perspectives affords researchers and teachers the opportunities to understand these perspectives and in turn make learning more accessible to the students (e.g., Clark, 1995; Finders, 1997; Johnston & Nicholls, 1995). Listening to student voices also allows researchers to understand the perspectives of students who experience the learning environment first hand (cf. Erickson, 1986). For example, Lincoln (1995) had argued that it is pertinent that students are provided an opportunity to voice the different aspects of learning that are important and meaningful to them so that researchers can shape learning environments that afford those learning opportunities to students. Others have argued that because students are "differently knowledgeable" (Cook-Sather, 2002; p.3) than adults about the use of technology, discourse practices and the use of education to their lives, it is important that students' voices be documented in relation to their perspectives on any designed learning environment (cf. Cook-Sather, 2002).

Inclusion of students in educational decision-making is also important from the perspective of promoting student learning. There is significant research that discusses that learners learn through meaningful interactions with different elements (e.g., people, language, objects) of the learning environment (cf. McDermott, 1996; cf. Bell et al., 2013). We need to include student voices in our designs in an effort to understand contexts and elements that are meaningful for students. Including student voices therefore serves the purpose of articulating learner needs and insights.

Rodgers (2006) used the term “descriptive feedback” to describe student perspectives on their own learning. She argues that descriptive feedback affords students the opportunity to voice their experiences in which the experience does not count as a “right” or a “wrong” experience. She argues that the lack of an evaluation of the learning experience could in turn facilitate learners to share experiences without worrying about their experience being evaluated based on others experiences (Rodgers, 2006).

In the 4R Activity, documenting students’ perspectives in relation to their own learning served two purposes. First, it served as an information gathering tool regarding what students considered as learning from engaging in the learning environment, along with what aspects of the learning environment helped or limited their learning experiences (cf. Rodgers, 2006). Second, it provided the researcher with valuable information on the different types of learning that students described as associated with the learning environment. This information could inform successive iterations of the learning environment to promote and support various types of learning for students in the future.

b. Connections between in-school and out-of-school learning. Research suggests that making education relevant to students’ life increase both engagement and learning (e.g., Lepper, Henderlong, 2000; Bransford, Brown & Cocking, 1999). When students draw connections between in and out-of-school contexts, their performance often improves (Hulleman & Harackiewicz, 2009). For example, a study by Hulleman and Harackiewicz

(2009) showed that an intervention designed to draw connections between students' lives and the science curriculum resulted in the enhancement of interest in science and classroom performance, particularly for "at-risk" students.

Despite the importance of drawing connections between the curriculum and the students' lives, scholars in education research (see Resnick, 1986; 1987), including scholars in science education (see Lyon, 2006; Bell et al., 2013) have highlighted the lack of connections that exist between the science classroom and the lives of the students. Three decades ago, Resnick (1986; 1987) had discussed the discontinuity in learning occurring in school and in out-of-school contexts. As one example of this disconnect she had discussed how arithmetic lessons in school were focused on acquiring formal rules of calculation and did not provide an avenue for students to draw on their knowledge of numbers acquired in informal settings (1987). Two decades since Resnick's (1986, 1987) work, Lyons' (2006) work with high school science students in three countries documented a "recurring theme" (Lyons, 2006, p.595) among students about the lack of connections between the science curriculum and their lives. In the study, students reported that science learning in the classroom made minimal connections to their lives outside school or to their personal interests (Lyons, 2006). Additionally, it was observed that high school students had "poor attitudes to school science relative to other subjects" (Lyons, 2006; p. 599). In the study, students reported that science learning in the classroom made minimal connections to their lives outside school or to their personal interests (Lyons, 2006). This caused them to have "poor attitudes to school science relative to other subjects" (Lyons, 2006; p. 599).The research further noted that many students who reported to be disinterested in school science were not necessarily disinterested in science outside of school (Lyons, 2006). Similarly, Bell and colleagues (Bell et al., 2013) also showcased cases that exemplified the lack of connections between school science and the lives of students. In the study, Bell and colleagues (Bell et al., 2013) highlighted accounts of students who were adept at carrying out sophisticated practices in out-of-school contexts that could be related to work in the sciences and in engineering (e.g., mixing perfumes, designing and constructing projects). However, these same students remained disinterested or were perceived as being disinterested in science learning in the classroom.

As noted above, research of different scholars spanning several decades has continued to highlight the lack of existing connections between the curriculum and the lives of students. Moreover, research has argued for the need for the research community to understand the importance of in and out-of-school connections and to support students in building connections across these contexts (see NRC, 2012; Lyons, 2006; Bell et.al, 2013). For example, Bell and colleagues argue for the need of the research community to "discover and then support the successful learning pathways of youth across social settings and developmental time" (Bell et al., 2013; p. 120) to develop the interests of students in school science and nurture expertise building in science and engineering disciplines (Bell et al., 2013).

Responding to the need to support students in building connections across settings, this study was designed to provide students with an avenue to voice their perspectives on the connections they see between the learning environment (i.e., the 4R Activity) and its applicability to their lives. However, it has been noted in research (e.g., Bricker, 2008) that it might not be sufficient to provide connections for students by simply embedding contexts meaningful to the learners within the curriculum. Bricker (2008) discussed that learners may not draw connections between the learning in the designed environment and relationships to their lives, or connections to what they consider meaningful. Thus to support the building of relevant connections between the learning environment and the students' lives this study argues that it is important to record students' accounts on the types of connections that students see between the learning environment and their lives. Once different types of connections have been documented, effective ways to support students in strengthening the connections and making other connections that might be relevant to their lives can be established in future iterations of the learning environment. Additionally, by documenting the reasons for why students do not see connections between the learning environment and their lives is valuable information that can be used to improve the designed learning environment in successive iterations. In an effort to understand the types of connections that students make between the 4R Activity in the science classroom and its applicability to their personal life, this study used students' narratives to explicitly document the different types of connections that students drew between the 4R Activity and its applicability to their academic and non- academic lives.

In addition to students' accounts providing valuable information for design improvements, helping students draw connections between in-school and out-of-school contexts also has been shown to promote learning. Effectively leveraging students' knowledge, experiences and community practices in the school setting can be a powerful aid in learning (Lee, 1993, 1995, 2006; Rosebery, 2005; Warren, Ballenger, Ogonowski, Rosebery, Hudicourt-Barnes, 2001). For example, Lee's work (1993; 1995; 2006) in

English Language literacy with African American youth discussed how incorporating activities and knowledge, which draw on students' social knowledge and practices, helped in promoting academic learning. Research in science classrooms has shown that students are able to employ their everyday experience productively as a resource for understanding science (Balenger, 1997; Rosebery, 2005). Furthermore, some researchers (Rosebery, 2005; Warren, Ballenger, Ogonowski, Rosebery, Hudicourt-Barnes, 2001) have also documented greater academic success in underrepresented student groups (e.g. non-white racial/ethic groups, language learners) when experiences and social knowledge were leveraged in science classrooms.

In spite of research suggesting the need to bring students' experiences, social knowledge and practices to the classroom, students in science classroomclassrooms are not afforded these opportunities often (Warren et al., 2001; Belenger, 2003). Responding to this, researchers have argued that science education should make consorted efforts to leverage students' social knowledge, experiences and community practices (Warren et al., 2001, Nasir et al., 2006). These researchers recommend that in leveraging students' ideas, knowledge and practices, science education should work towards establishing continuities between students' experiences and practices and scientific ways of knowing (Warren et al., 2001;Bellenger, 2003).

In order to assist educators in effectively using students' social knowledge and practices in the classroom, Nasir and colleagues (2006) proposed three aspects for the design of such activities. First, they recommended making visible the specific practices in the activity and the purpose of the practices in the academic discipline. Second, they recommended active engagement of students in disciplinary practices so as to allow students' to identify with the practices of the academic discipline. Third, they called for inclusion of "occasions for metalevel analysis" (Nasir et al., 2006, p. 497) that help students' see the connections between everyday practices and the academic practices. Nasir and colleagues (2006) discussed that these three aspects in a learning environment facilitate learners in bringing their knowledge and practices rooted in everyday experiences to the classroom. Additionally, it also restructures the relationship between the student and the educator/teacher because students often draw on social knowledge and practices in which they might have greater expertise than the teacher. The restructuring of the relationship allows both teachers and students to take on the role of an expert and a novice (Nasir et al., 2006). This in turn is argued to facilitate students in more closely identifying with the classroom activity and a having a greater sense of belonging to the classroom community (Nasir et al., 2006).

The 4R Activity together with students' interviews reflected some of the aspects proposed by Nasir and colleagues (2006) for the design of activities that leverage students' ideas and everyday experiences. For example, students were introduced to the 4R Activity through an informational video explaining how scientists, engineers and other professionals engage in processes similar to the 4R Activity during knowledge construction (see Chapter 2 & 3, this dissertation). Secondly, the participation in the 4R Activity allowed students to actively engage in the practices of the scientific discipline. Lastly, the interview conducted with students (reported in this chapter) functioned as an opportunity for students to engage in "metalevel analysis" (Nasir et al., 2006, p.497) and discuss the connections they saw between the different stages of the 4R Activity and their everyday experiences, knowledge and practices.

c. Connections between classroom activities and how they relate to the work of professionals. The *Framework* (NRC, 2012) and the NGSS (NGSS Lead States, 2013) emphasize the learning of science in a manner that reflects how science is practiced in the real world. For example, the *Framework* (NRC, 2012) discusses science learning that highlights the social nature of science where knowledge building is collaborative and takes place in a social context with well-developed norms (NRC, 2012). In an effort to support collaborative knowledge building, the 4R Activity was designed to apprentice students in the formal activity of peer review undertaken by professional scientists and engineers when building scientific knowledge.

With respect to engaging learners in authentic activities, Brown, Collins and Duguid (1989) discussed that as learners begin to engage as novice practitioners in authentic activities (such as the 4R Activity in this study), they build a rich understanding of the activity as well as the application of the activity in other contexts. This study argues that it is important to document the connections students make between the activity and the work of professional scientists and engineers, as well as to professionals in non-science careers. Documenting the connections that students draw between the science activities and the work of professionals in both scientific and non-scientific disciplines is a step towards making science relevant for all students (NRC, 2012). The *Framework* (NRC, 2012) clearly outlines that one of the goals of science education is to equip students with skills to pursue a career of their choice that includes but is not limited to careers in science, engineering and technology (NRC, 2012). This study argues that by documenting student perspectives, the research community can support students in building successful connections between the classroom activities they engage in and how those activities translate to the activities undertaken by professionals.

Methods

This section presents information on sampling procedures, the data sources used and the analysis employed to address the research questions.

Sampling Procedures

Five teachers located in four different school sites implemented the curriculum along with the 4R Activity in their classrooms. The schools were located in New England, West North Central and East North Central divisions of the United States. The participants were considered in the sample only if the student had parental/guardian consent as well as his/her own assent for the author of this paper to use his/her interview data. Additionally, sampling selection was based on students having participated in all three 4R Activities. In total 173 students from a total of 230 students satisfied the above two criteria (also see Chapter 2 & 3, this dissertation).

Out of the 173 participants, 22 students were purposefully selected for this study. The 22 students were selected from across the classrooms of the five teachers. Roughly, one-tenth of the students from each class were part of the 22 students sample. The selection of students was done to attain maximum variation within sampling (Maxwell, 2013) across the three 4R Activities with respect to the demonstration of varied levels of growth in the quality of the peer review, reflection on peer review and the incorporation of feedback during revision. Students with Individualized Education Programs (IEPs) and students classified as English Language Learners (ELLs) were purposefully included in the sample to have a more inclusive understanding of student experiences.

The 22 selected participants consisted of ten females and twelve males. This distribution was representative of the gender distribution (45% females) in the larger 173 students sample (see Chapter 3, this dissertation). Additionally in the sample used in this study, two students classified as English Language Learners (ELL) and two with an Individualized Education Program (IEP) were also included. The inclusion of the ELLs and students with IEPs were also representative of their distribution in the larger sample (12% ELLs; 9% IEPs) of students that participated in the 4R Activity (see Chapter 3, this dissertation). Participant information is presented in Table 4.1.

Table 4.1
Participant Information and school location

School ID	U.S. Census Bureau Designated Divisions	% Enrollment in Free/Reduced Lunch Program	Participating Teacher*	Grade	Students with complete data	Number of students interviewed	Number of interviewed students classified as ELLs	Number of interviewed students with IEPs
B	New England	36	Ms. Lata	6	17	2	1	0
B	New England	36	Ms. Asha	6	17	2	0	0
C	West North Central	30	Ms. Priya	7	33	5	0	1
D	East North Central	75	Ms. Laila	7	30	4	0	1
E	East North Central	12	Mr. Amar	6,7,8	76	9	1	0
				Total	173	22	2	2

*Pseudonyms used for teacher names

Data Sources

Semi-structured interviews (Weiss, 1994) were conducted with the participants and served as the primary source of data. The questions asked in the interview were crafted based on the research questions guiding this study (refer to Appendix F for the interview protocol). The interview protocol was refined after piloting with three students from a participating classroom. In an effort to not misrepresent the participants' words, elaboration questions were asked within the interview to help in clarifying what the participants meant in certain instances. Additionally the participants were informed that the purpose of the interview was to gather their perspectives to inform design decisions.

The semi-structured interviews were audio-recorded and transcribed. Given the geographical dispersion of the school sites, the interviews were conducted face-to-face or through video conferencing using Skype ™ software. This was similar to the method of interviewing adopted by other researchers with geographically dispersed population (Sedgwick and Spiers, 2009; Hanna, 2012; Deakin, & Wakefield, 2013). Research comparing the quality of responses from face-to-face interviewing and online synchronous interviewing (e.g., Skype) has shown to produce similar results with the occurrence of pauses and repetitions not differing significantly in the two formats (Denscombe, 2003; Cabaroglu, Basaran & Roberts, 2010; Weinmann, Thomas, Brilmayer, Heinrich, & Radon, 2012).

Analysis

The semi-structured interview protocol was designed to capture student experiences pertaining to three exploratory categories, namely, (a) comparison between the peer review, reflection and revision process in the 4R Activity with respect to that undertaken by scientists, engineers and other professionals; (b) student perceptions of learning from participating in the 4R Activity; and (c) student ideas on the applicability of the 4R Activity to academic and non-academic contexts. The data from the semi-structured interview was first open coded and then focused coding was developed. This was followed by integration into conceptual categories and themes that resulted in claims grounded in the data (Emerson, Fretz & Shaw, 2011).

Some of the validity strategies that were employed were analyzing the data in iterative cycles where the data corpus was reviewed repeatedly to test the validity of the claims that were being generated from the data (Erickson, 1986). When coding data and providing it as evidence for a category, attempts were made to ensure that the interpretations stayed as close as possible to research participants' own words. This was done to preserve the local meaning, as defined from the viewpoint of the participant (Erickson, 1986). Additionally triangulation in the form of environmental triangulation (Guion, Diehl & McDonald, 2011) was employed given the data was collected from different school sites and from the classrooms of different teachers.

Results

The results in this section are presented in three sections, each addressing one of the three research questions. Pseudonyms in the format of “Student Number” are used for students when presenting student responses. Additionally the location of the students, as well as information about their gender, is not presented to prevent stereotyping of students’ responses based on location or gender. For students classified as ELLs or students with IEPs, this information is only revealed when a pattern is seen in their responses that informs our understanding of designing learning environments to support all learners. The author of this paper has concealed this information in instances when it seemed that revealing the information might lead in stereotyping the learning abilities of students classified as ELLs or students with IEPs.

Learning from Participation in the 4R Activity

The research question that drove the analysis was as follows:
What are student perspectives on learning from participating in the 4R Activity?

In the interview the participants were asked about their perspectives pertaining to learning from participation in the 4R Activity. They were also informed that if they felt that they did not learn anything from their participation, they could express that too. None of the twenty-two students reported that they had not learned anything from their participation. However no pattern emerged between students' reports on learning and their demonstrated learning (e.g., improvement overtime in providing peer review, incorporating feedback during revision) for which they had been sampled. For example, Student 16 had demonstrated improvements over time in incorporating feedback during revision and had been selected into the sample based on that criteria but Student 16 reported that learning how to provide peer reviews was the most important aspect of learning from participating in the 4R Activity. Student accounts generated eight themes with respect to learning associated with the 4R Activity. Figure 4.1 provides the different aspects as reported by the students.

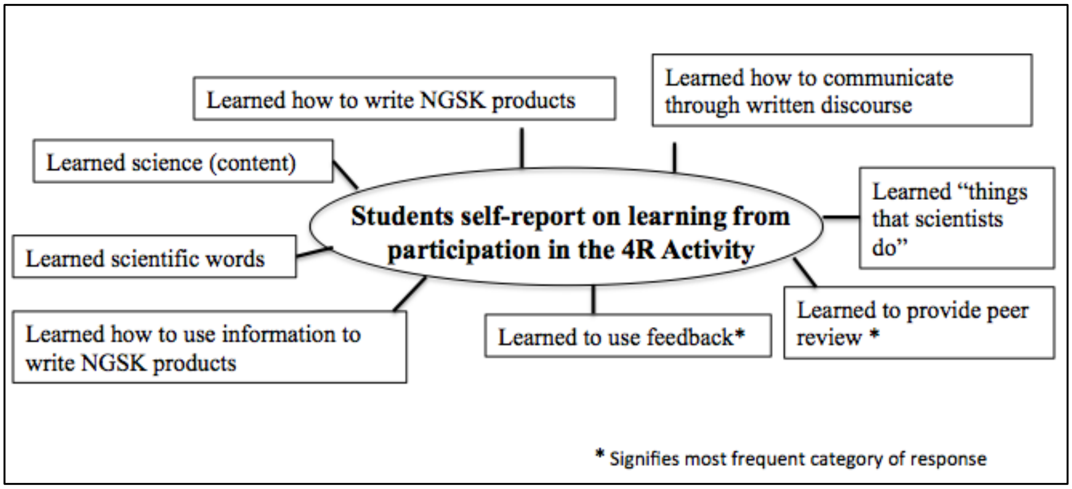


Figure 4.1. Students' perspectives on learning associated with the 4R Activity.

Student accounts on learning to provide peer reviews. The most frequent perspective on learning associated with participation in the 4R Activity was that students *learned to provide peer reviews*. Out of the 22 students interviewed, 35% reported that they *learned how to provide peer reviews* from the 4R Activity. An example student comment about this is provided below:

> Student 1: "I think I learned how to become better at peer reviews. I think I got better and I could see it. By the last one [4R Activity] I was more careful and I picked up things that I wouldn't have otherwise…"

Student 1 further elaborated on how he/she learned to provide peer reviews.

> "I think I learned that [to provide peer reviews] from when I read the scientists review carefully and tried to do similar stuff. And I know I did well because the scientist praised me in the feedback and said that I was giving good peer review and then told me what I should keep in mind for next time."

The above quote from Student 1 describes the process from the *Reflect* stage of the 4R Activity where students reflected on the peer review they provided in light of the scientist's review and the scientist's feedback on their reviewing skills. Apart from the students who provided the elaboration by themselves, such as Student 1, the other students were specifically asked to elaborate on what had helped them in learning how to provide peer reviews. Their responses also revealed that reading the feedback that the scientist gave on the peer reviewing skills was helpful in learning the peer review process.

> Student 16: “Scientist feedback was helpful because it was -- I don’t know how to explain it but it kind of felt like [pause] like it was easier to improve on how to give review when I did the reflection. I remembered things that the scientist said and how he did it and I tried to use it wherever it applied. Sometimes it didn’t apply to what I was reviewing but other times it did and I remembered.”

Additionally students also reported that having an opportunity to read the scientist’s review helped them in improving their peer review skills.

> Student 5: “Reading the scientist’s review was good, because I could be like, umm, like look at things that they say and how they say it and then I could make a note of it and record it in my reflection. It was helpful. I think in my reflections I tried to remember what the scientists said. Also before the activity, I would reopen the old ones [reflection activity] and just glance at my reflection to -- you know, remind myself.”

Students provided a variety of responses when asked about how providing peer review contributed to their learning. Their responses could be categorized into five themes, namely, (a) helpful in future writing of the NGSK products; (b) provided a different perspective on how to revise their own NGSK product; (c) gave an opportunity to help peers improve; and (d) gave an opportunity to see how scientific words were used by the peer in the NGSK product, and (e) would be helpful in professional life. Representative quotes for each of these themes are provided in Table 4.2.

Table 4.2

Aspects of peer review that students found beneficial

Emerging Themes	Exemplar Quotes from Students
Helped in future writing of the NGSK products.	Student 11: "It helped me write my own and it helped me keep those things in mind, like when I was reviewing somebody else's work I come across things that I were like "oh! maybe I should include this in my future prediction or explanation too"."
Provided a different perspective on how to revise his/her own NGSK product.	Student 5: "I think it gave a different perspective on how to write your own. Don't get me wrong, I think it did -- my 4Rs did improve one after another because I had done like the revision but I think they also improved because I got to review others work and learn from that. So I think the peer review thing is good."
Gave an opportunity to help his/her peer to improve.	Student 13: "Yeah I got to see what other people wrote and got a chance to help them improve it."
Gave an opportunity to see how the peer used scientific words in the NGSK product.	Student 9: "I saw how they used specific words and that how some things were said by my peer so my own explanation or prediction whatever it in that activity was, got better, I could improve it."
Helpful in professional life in the future.	Student 1: "That [giving peer review] would be very helpful to me later on because I want to become a teacher or a professor I haven't decided on that right now, but definitely one of those things. And learning how to give feedback is what they have to do all the time, so I feel that I learned it now already and now I know that you have to really read carefully and think of how to help the other person, instead of just trying to not take it [giving feedback] seriously and finish it quickly."

Students' perspectives on the double-blinded peer review process. Given that the double-blinded feature was a feature that was unique to the 4R Activity, all students were asked about their perspectives on the double-blinded review process. When prompted about the double-blind review process, data revealed that no students had previously engaged in the double-blinded review process in school science or in other classes. 90% of the female students and 20% of the male students reported that they preferred the double-blinded review process and cited reasons for their preference (see Figure 4.2). With respect to students who reported that they did not prefer the blinded process, the most frequently cited reason was that the students did not want to be “harsh” to the author of the NGSK product. Additionally, some students also felt that if they knew the identity of the author they could offer encouragement to the author if they had prior knowledge of the challenges that the author faced in the science classroom.

> Student 15: “I thought I would have liked to know so I didn’t have to judge that person only on how well they did. Like maybe that person needs extra help and is not so good in science and then you could not be harsh with them…some people just don’t like science or are not good at it and you want to say good job to them so they try harder.”

Out of the 80% of males who had reported that they did not prefer the blinded process, 60% of them indicated that the blinded process might be helpful for other students who otherwise could have given a biased review.

> Student 12: “It didn’t matter to me because I would have done stuff the way I wanted to anyway. But it might help others in reviewing because for some people if we knew who the person was they would be like okay I don’t really like this person so I’ll just give them a bad review no matter what it is.”

The themes that emerged from students’ reports on the double-blinded review process are shown in Figure 4.2. Given that students who did not have a personal preference to the blinded process had indicated that blinding might be beneficial for others, their accounts were included in preference toward the blinded process. In other words, having a preference toward the double-blinded review process encapsulated both personal as well as preferences for its use by others in the class.

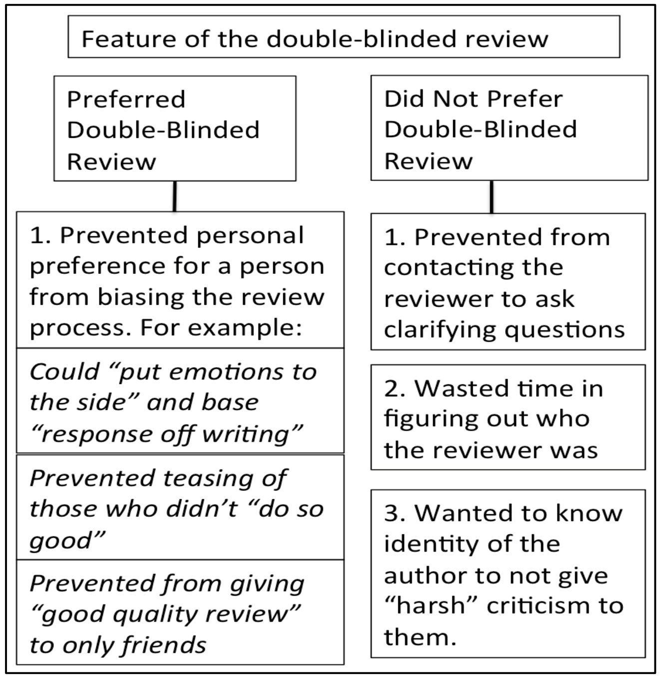


Figure 4.2. Aspects of providing blinded peer review that the students found beneficial.

Student accounts on learning to use feedback. Another frequent theme related to learning associated with the 4R Activity was *learning how to use feedback*. 27% of students reported that they learned how to use feedback from the 4R Activity.

> Student 18: "I learned that, how to read the feedback and then use it to revise my explanation and prediction…. Reading peer and scientist feedback carefully and then using it, that was the most important thing I learned."

Some students reported that the feedback helped "make more sense of science" to them, which in turn allowed them to write the NGSK product so as to make it more "understandable to other people" as well.

> Student 19: "It helped me make more sense of the science and I felt that since I got help that my explanation and prediction would make more sense to my mind and it would get me to understand the science more often. If I could make more sense of it then I could write it to make it more understandable to other people also."

All students reported that feedback helped make their revised NGSK product stronger than their original NGSK product. A student with an IEP described that the revised NGSK product was stronger when assistance by the teacher was provided in how to incorporate the feedback.

> Student 22: "I felt like it was stronger with more help and advice... and I felt like that when I got help from [Teacher's Name] then it was stronger and it was only sometimes stronger when I did it on my own."

Students described that they thought that their revised NGSK product was stronger because, (a) it was proof read and did not have spelling and typographical errors, (b) because peer review had pointed out inaccuracies that were then "fixed" in the revised product, and (c) because the revised NGSK product had more scientific terminology incorporated into it.

Other perspectives on learning associated with the 4R Activity. Besides the two most frequently emerging themes discussed above, the other themes that emerged from student accounts of learning from participating in the 4R Activity were, (a) learning of science content, (b) learning how to write NGSK products, (c) learning to use information in writing NGSK products, and (d) learning to use scientific terminology. Students reported learning science "content", and provided examples of the "content"(e.g. greenhouse gases, information on a particular species). Additionally some students also reported that the scientific questions were based on real life events and that helped make them "interesting".

> Student 8: "Mostly I learned the content… I thought that the scientific questions were interesting…I thought that part of what made them interesting was that they were based on real things. An example would be that I learned about greenhouse gases and how they can affect the climate, also how that affects the species."

A student also reported to have learned how to use the information provided in constructing their NGSK product. When asked to elaborate on their responses, Student 22 highlighted the difficulty experienced initially in using the information provided to write a NGSK product as the information was presented differently than in the lessons.

> Student 22: "At the beginning I had a hard time with them [pause] the using the information. It was different from the lessons because in the 4R all the information was provided at one time but in the lesson it is from all over and then you use all of it to write your explanation but here it was lots of things in one place…"

However students reported that over time, they learned how to use the information provided to construct the NGSK product.

> Student 3: "But by the end it kind of became second nature like when you were reading the information that was given usually we had a graph and then a few sentence thing. I would already be thinking about okay I can use this as evidence, I can use this as reasoning. So by the end it kind of became second nature to be able to do that. That would be maybe the main thing I would think of for learning from the 4R."

Two ELLs (who attended different schools) described that the most important learning from the 4R Activity was learning the use of scientific terminology to make the NGSK products more comprehendible to the audience.

> Student 9: "The most important thing was the scientific words that the feedback told me to use. I learned the most about how I need to use them and not just saying things like over and over again. Because those words [scientific terminology] get to the thing more fast and clear and everyone gets the words and know what you are trying to say."

For some students learning how to communicate through written discourse was the most important take-away from the 4R Activity. Students most often cited the process of constructing and revising NGSK products using feedback as contributing to that learning.

> Student 10: “Well writing and revising kind of helped me become more better at explaining to people my explanation, without talking and telling them, so by only writing. That was different from how it’s in the class. It was hard at first but the feedback really helped me to get better at telling what I wanted to say in writing. I think that was the main thing that I learned, on how to talk through writing”

Moreover, two students reported to have learned multiple aspects from participating in the 4R Activity, which they termed as “things that scientists do”.

> Student 20: “I learned more things that scientists do, like write clearly for other people, to read and receive review, and revise, and also give reviews. Oh also reflect on how to be better on things like writing, the science, and giving better reviews… I liked it because it will help me become a better scientist in the future.”

Similar to Student 20 who reported to have “learned things that scientists do” and expressed the aspiration of wanting to be a scientist, many other students also reported learning aspects that they thought would be relevant to their career or future aspirations. Some examples provided by students highlighted how they linked the 4R Activity to practices that might be useful to them in their careers. For example, Student 19 had earlier expressed how the feedback in the 4R Activity helped him/her in “making sense of science”, which Student 19 could then use to communicate more effectively to an audience. Student 19 then provided further elaboration by providing an example of the importance of clearly articulating and communicating knowledge with an audience in professional settings. Additionally, Student 19 also highlighted the importance of learning how to receive and provide feedback in the workplace.

> Student 19: “Because teamwork is so often--, I don’t know how to say it but it is important and so I think once you’re older you’re going to be working with other people and you have to learn how to clearly tell them what you know and to get advice from them and you have to learn how to take and give constructive criticism to other people.”

Other students provided specific examples of linking the learning of the 4R Activity to a specific career. For example a student reported that providing feedback was the most important learning from the 4R Activity and then linked it to the career aspiration of wanting to be a teacher. Another student reported the use of feedback as what he/she had learned from the 4R Activity and tied it to college preparedness stating that using feedback and revising was prevalent in college classes. Another student, who wants to be an engineer, also reported that the use of feedback and revision would be helpful in engineering design tasks. Examples of student responses that highlight learning tied to future aspirations are provided in Appendix G.

Applicability of the 4R Activity in and out of school contexts

The research question addressed by this analysis was:
What applicability do students see for the 4R Activity in science class, other classes and/or in their life outside school?

a. How do they describe this applicability?

Students were asked if they had done activities similar to the 4R Activity in the science classroom or in other classes before. They were informed that they could mention activities that utilized one or more of the four stages of the 4R Activity (i.e., respond, revise, review or reflect). 90% of the students reported to have used two 4R Activity stages in English Language Arts (ELA). For example, a large majority of these students reported using review and revision when writing essays. On the other hand only 41% of the students reported that they had used activities similar to the 4R Activity in the science classroom before. Out of the 41% of students a majority of them (35%) reported having done peer review in the science class before. Most instances of using peer review involved science projects where students provided reviews in pairs or as part of a group. When asked for examples of use of the peer review process in science projects, the students reported that reviews functioned primarily as an evaluation and the receiver of the review was not expected to revise. Only two students reported that review in their science project was used to improve the project.

All students were asked if they could see any benefit in using the 4R Activity in science classes in the future and all the students responded that they saw benefit in using the 4R Activity in the science classroom. Some students provided examples of scenarios of how the 4R Activity could be used in the science classroom. For example Student 13 described how the 4R Activity could be used for difficult concepts in science because in Student 13's experience the feedback and the revision process was helpful in understanding "confusing" and difficult science concepts.

> Student 13: "I think working on structure, bone structure and our muscles and things like that because those can be confusing. And the 4R Activity helped me to, to understand confusing things better because I got feedback and then I could revise. So it was not just one-shot thing where you do it and then that's it. It was okay to get feedback and make it better and it did not mean that you did not understand anything. Feedback means that you can make it even better. So we can use it [the 4R Activity] for all confusing things...like the difficult stuff where you need help and you get time to work on things more than once."

Students also described how activities similar to the 4R Activity could be used in other classes, such as social studies (when writing historical accounts) and in mathematics (when providing an explanation on how he/she arrived at the answer).

> Student 19: "And I think it would work in social studies and math. We usually in social studies have to write papers about, um -- historical times. And that would be easy if we wrote it, and we shared it with a peer, and they got to tell us, "This is what you need. This is what you work on." And then you go back again and then you fix it. It would work in math because you need to work with partners to understand math. And you, sometimes, do have to write down your explanation on how you got this. And yes, you're going to need a partner to tell you. "I think

> this is wrong or I think this is right. You could add a little bit more of this in the details." And then we could go back and change it."

Similar to providing examples of the use of the 4R Activity in different contexts, some students described the 4R Activity as a "tool" that could be used in multiple contexts within the school setting.

> Student 19: "One of the points I wanted to make about the 4R activity was that it didn't seem to pertain specifically to anything, like a particular topic. I mean it did but you could easily use it in other places in science and in other places. Like you learn something else like chemistry and then you use the 4R Activity as a tool to tackle a question, I mean that is how I think it can be used. It is not a lesson like other lessons but it is like a tool."

Additionally many students provided examples of how they would use the 4R Activity in out-of-school contexts, such as activities they pursue for leisure. Students provided examples of using the 4R Activity in non-academic settings such as sports, music, dance, fine art, and video gaming. Table 4.3 provides the activities and some examples provided by students.

Table 4.3
Examples of using the 4R Activity in non-academic settings

Using the 4R Activity in non-academic settings	Student Examples (Direct Quotes)
Sports (e.g., Basketball)	Student 21: "basketball is usually jump shots -- so getting feedback from other people and then going back and practicing it over and over, like the 4R activity did with scientific explanations and going back and seeing what you know and how much another person could help in basketball with either layups or jump shot shooting. Yeah definitely."
Video Gaming	Student 7: " Yeah because you usually like with pretty much any sport you watch yourself play. You respond to questions like how could we do this differently and then you would review your teams moves and yeah use the 4Rs to varying degrees. Like by yourself or with your team think about it and use it and then redo it."
Fine Art	Student 15: "Drawing: Oh yeah like in reviewing and reflecting -- someone above you could review yours and you could review theirs and take information from them and they could maybe take

	information from you. And then for reflecting you could reflect on someone else's art work from in the past."
Dance (e.g., Ballet)	Student 13: "Well we could do our routine and the other people would watch and they could help us or give us examples of ways we could improve our technique."
Music (e.g., Playing the saxophone)	Student 12: "I normally realize that I might have messed up on stuff and I might have wrote the wrong note names and I go back and revise that to make it better…I do it on my own sometimes, and sometimes my parents hear me and they're like that doesn't sound right, maybe you should double check. I can use the 4Rs in umm…[pause] I can revise mine, and I can review it. I don't know how I would reflect on that. Maybe next time taking more time to write in the note names, but I think that's a way.

Additionally, one student from a school site that had adapted the 4R Activity to the ELA class described the 4R Activity as a *link* that could bridge disciplinary boundaries as well as provide continuity between academic and non-academic contexts.

> Student 15: " I think the 4R Activity can be used in so many things. I mean you could maybe use it in different classes like we did in science and ELA. It helps to link things because all classes are on different topics and we go from one class to other but we can use a process like the 4R Activity to link stuff up. Also we could use it in maybe at home or when we play and that can be one way we can link the things we do at home or with friends and stuff like that with class stuff."

Connections between the 4R Activity and the work of scientists and engineers

Given the 4R Activity was modeled to simulate the process of peer review undertaken by scientists and other professionals, it was important to explore students' ideas on whether they considered that scientists and other professionals engage in similar activities. The research question being addressed was as follows:

Do students draw connections between the 4R Activity and the activities undertaken by scientists, engineers and other professionals?

a. What similarities or differences do they describe if any?

All students in the sample were asked if they saw connections between engaging in the 4R Activity and the activities that scientists and engineers engaged in. 90% of the students reported that they thought that the 4R Activity was similar to the practices of scientists and engineers. Students further elaborated by providing examples to how they viewed the 4R Activity similar to the practices undertaken by scientists and engineers. Students were asked how they knew that scientists and engineers engaged in these activities, 30% of students reported that engaging in the above activities seemed "probably what scientists and engineers do". 40% of the students reported that they had seen scientists and engineers engaging in these activities on science shows on television (e.g., Nova on PBS) and on the Internet (e.g., YouTube science channels such as NASA and SciShow). Some students also reported that they knew that scientists and engineers engage in these activities based on the 4R Activity video that students were shown prior to the first 4R Activity. The video mentioned that scientists and engineers engage in the activities of peer review and revision. The remaining 30% reported to know a scientist or an engineer and cited that as the source of information about scientists and engineers engaging in the activities of receiving and providing peer review and undertaking revision.

Many students provided examples of how scientists and engineers might engage in the activities of providing peer review, incorporating feedback and reflection and elaborated on why it might be beneficial to their work. Many of the examples provided by students highlighted that receiving review and revising was necessary in the work of scientists and engineers as it lead to the refinement of the product or work that had been review.

> Student 14: "Like for engineers when they're working on something and they needed help they could ask someone and they could revise over it and help them to make the design better. I think it would be very helpful so when they're working on vehicles or when building engines and everything that they get help from other people who can look at things that you might have missed so they can be safe and everything goes in the right spot."

Additionally, some examples also discussed how science is a collective effort and receiving review allowed for others from the community to contribute to knowledge building before it was available to members from outside the community. Additionally receiving peer review prevented the work of science from being biased.

> Student 8: "Yeah so getting a review and revise are important because science is teamwork. In science it sort of makes sure that it is like not just an individual's thoughts and it's not biased and all of that. Like some experts and people who do the work have looked at it before [pause] the public can see it."

Student examples on providing peer review highlighted that peer review allowed the person to be aware of the work done by their colleagues, which in turn could influence their own work.

> Student 5: "Well because sometimes scientists might be working on something and not realize stuff that is being done by other scientist people. Sometimes you need to look at the work of other people to notice other things that you might not see because you're so focused on your own stuff…You could review other peoples work and that way scientists could learn new stuff and maybe use it in their own stuff"

Some other examples provided by students on the similarities between the 4R Activity and the work of scientists and engineers are shown in Table 4.4.

Table 4.4

Similarities between the 4R Activity and the practices of scientists and engineers

Practices of Scientists and Engineers	Examples of similarities between the 4R Activity and the practices of scientists and engineers (Direct Quotes)
Providing Review	Student 7: "The review process was like how engineers do it. Yeah because my dad's an engineer and everybody's always giving him their stuff and he has to review it and give suggestions..."
Receiving Review and Undertaking Revision	Student 2: "It is similar to the 4R Activity in how we did review and revise. I'm guessing that a scientist would ask another scientist to review if they've got a -- I think they call it a article? If they've got a article they might ask another scientist to review it so they can revise it…that's how we did it too."
Reflection	Student 17: "Engineers they try to make technology better and better. They reflect on the technology to make it better next time. That's like the reflect where we were thinking of how to make giving reviews better and better by asking "what have I learned from doing the review"."

Out of the 22 students, 10% of students (n=2) reported that the practices in the 4R Activity differed from the activities of professional scientists and engineers. When inquiring about how the students knew about the differences between professional settings and the 4R Activity both students reported that their parents were scientists and they had gotten the information from them through discussions and/or observations of how their parents worked.

One student stated that the practice of peer reviewing and revision undertaken by scientists and engineers differed from the 4R Activity with respect to how scientists and engineers engaged with them. For example, the student reported that scientists and engineers were not provided a question to respond to, like in the 4R Activity but they "look for the question they want to ask" and therefore the scientists and engineers providing reviews may not have answered an identical question.

> Student 6: “ I think that scientists and engineers look for questions they want to ask and its not like how we were given a question in the 4R Activity that we had to make an explanation or a prediction. So when they [scientists or engineers] get reviews, those scientists or engineers [giving the review] might not have answered the same exact question or build the same exact thing but maybe similar questions or on topics they know.”

Additionally, another student reported that the time duration of receiving and providing reviews differed in professional settings as compared to the 4R Activity and that scientists and engineers may not be providing reviews and receiving reviews at the same time and therefore felt that in that aspect the 4R Activity was different than what scientists and engineers engage with.

> Student 5: “ The 4R Activity is different than how scientists and engineers do these things because they do not give reviews and get them at the same time, definitely not in one day. They get a lot more time to do that. Sometimes you get reviews and sometimes you give reviews, but you definitely don’t do it together.”

Additionally Student 5 also reported that the reflection process of scientists differed from the 4R Activity because scientists already knew how to review and might not need to reflect on their peer reviewing skills. Moreover Student 5 added that scientists might engage in reflection of other types such as on “how to improve the solution” of the problem they are investigating.

Connections between the 4R Activity and professionals from non-science domains

To draw parallels between the 4R Activity and the activities of professionals in addition to scientists and engineers, students were asked about extra curricular activities they liked to pursue or things they did for leisure (e.g., dance, music, watching TV). Based on their interests, students were asked if they could see the professionals as engaging in activities like the 4R Activity. A majority of students (90%) said they could see professionals engaging in activities similar to the 4R Activity. Table 4.5 provides some examples given by students of professionals engaging activities similar to the 4R Activity.

Table 4.5

Examples of professionals using the 4R Activity

Professional	Student Examples of Use of the 4R Activity by Professionals (Direct Quotes)
Guitarist	Student 14: "I think they use the reviewing --they record themselves play and then they listen back on it to see if they like it or not and if they don't they change it. But it is like self-reviewing most of the time and some time it can be by others also."
Video Gamers	Student 2: "I suppose they could maybe if there is someone they know is sort of hard to beat in a tournament or something they could try to review their strategies for playing to try and see how they could beat that person or whatever."
Singing	Student 19: "Well, I think it could be useful for them when they want to write songs. And then they could have someone back it up and say, "I really like your songs." And that person could be more specific about what they liked and what should be improved. I think it could be useful."
Dancer	Student 18: "Yes a dancer would use it because they could get review on the moves and get feedback, which they could use to revise their moves."

Two students provided examples of how the 4R Activity differed from what professionals would engage in. One student reported that dancers could not use the 4R Activity because in the 4R Activity there was a prescribed format on writing NGSK products, which was different from dancers who could re-interpret the moves in a way that was different from other dancers. Additionally, in the 4R Activity, writing the NGSK product had to be based on evidence whereas in dance the moves could be based on personal opinion.

Interviewer: "What makes it difficult for a dancer to use the 4R Activity?"

> Student 22:"I said no because there is more detail in the 4Rs than there is to dancing. Because in the 4R you have to answer the questions a certain way that have to do with that subject and there is a certain way to write the explanation but with dancing you can just express yourself freely and do new things and not do exactly the moves that are done by others. There is more freedom in dance but in the 4R Activity you can't make up things and say " I will write this because I feel like it" it's more on what evidence you have."

Another student provided an example of how the feedback provided by the scientists in 4R Activity differed from practices of professionals who gave feedback to their mentees. The student used the example of feedback received by contestants on singing and cooking reality TV shows. Some of the differences highlighted by the student were that there existed a hierarchical relationship between the provider and the receiver of the feedback in the television shows. This was in contrast to the 4R Activity, where the student described that everyone was "on the same level". Another point of departure was that on competitive reality shows, an expectation that the receiver of the feedback (i.e., the contestant) had to implement all the feedback that was provided and did not have the agency to choose what part of the feedback the contestant wanted to incorporate. Whereas in the 4R Activity, the receiver of the feedback could decide what feedback they wanted to incorporate.

> Student 7: "Well I mean I think that the kind of cool thing about the 4R Activity is everyone is kind of equalized, you're all on the same level. With those [TV competitions] there's definitely someone on top and someone on bottom. And it's pretty clear who is on top, like the mentor; and feedback is coming one-way and then the person has to just do that no matter what. Here in the 4R Activity, the peer was someone from our class and the scientist was more knowledgeable than us but he [scientist] wasn't saying "you have to do it like I say", you know, it was more like "here, I think this will make it better" and you could read it and say... "hmm I don't agree with this thing but I like the other suggestion". '

Discussion

Multiple salient themes emerged from the data that are discussed in this section. The emergent themes have been further categorized into two categories to aid in the process of discussion. The first category is associated with the information gathered on students' perspectives of learning from the 4R Activity. This category helps in understanding how the information can help inform successive iterations of the 4R Activity. The second category pertains to insights based on the connections students drew between the 4R Activity and their life both in and outside of school. The second category has implications for how science educators could support student learning to make it relevant to the lives of the students.

1. Student perspectives on learning after engaging with the 4R Activity

Exploring students' perspectives on their own learning associated with the 4R Activity allowed this study to capture an aspect of learning that was not previously captured by other data on student learning and the use of the 4R Activity (see Chapter 2, Chapter 3, this dissertation). In this study, the learning accounts of students provided a range of aspects that students associated with participating in the 4R Activity. These accounts ranged from reports on learning that were tied closely to the 4R Activity context (e.g., learning the scientific concepts of climate change) to other, broader aspects, such as communication through written discourse, using feedback, providing feedback and engaging in "things that scientists do" (refer to Figure 4.1 for different types of learning). Of these accounts the most frequent learning cited by the students was, (a) learning how to provide reviews and, (b) learning how to incorporate feedback in revisions. Data on students' competency with these activities (see Chapter 3, this dissertation) had revealed that students were indeed progressing over time in the reviewing competency, as well as in revising their NGSK products (see Chapter 3, this dissertation). Through students' own accounts, this study could explore how students saw these activities as contributing to advancing their learning.

Students provided a variety of responses on how receiving and providing review helped in learning. In addition to building reviewing competency, students cited that having reviewed their peers' work helped inform the construction of their future NGSK products. These accounts are in line with research that discusses that peer reviewing benefits the reviewer's own writing (Lundstrom & Baker, 2009). Additionally, students also provided examples of how receiving reviews helped them in the understanding of the science concepts and helped them in making the NGSK product stronger with more scientifically accurate information, effective incorporation of scientific terminology, and having less grammatical and spelling errors. The student accounts on the benefits of receiving and providing reviews are also supported by research that has shown that students gain from engaging in receiving and providing feedback by scaffolding each other's learning (e.g., Lunstrom & Baker, 2009; Liu &Carless, 2012; Teo, 2006).

In the 4R Activity the process of receiving and providing reviews was undertaken through a double-blinded review process. Additionally, students were supported in the practice of providing reviews through engaging in a reflection activity. Discussing students' perspective on these two aspects is important because the double-blinded review process was a novel feature of the 4R Activity while the reflection activity was added to support students in developing their reviewing skills. Students' feedback on these two aspects provides a first- hand account of how students perceived the role of these two aspects in supporting their learning.

a. The Double-blinded review process

Students most commonly cited receiving and/or providing reviews as aspects that they learned from participating the 4R Activity. Students reported that the blinded process allowed them to provide unbiased feedback based solely on the quality of the submitted NGSK product. The students who reported to not prefer the blinded process citied the inability to ask clarifying questions to the reviewer, given the identity of the reviewer was concealed. Given these perspectives, it is clear that the double-blinded process can be used to provide feedback, but the system needs to be refined to allow for authors to ask clarifying questions to the reviewers.

Other students who did not prefer the blinded process reported that the anonymity of the process prevented them from providing feedback that was sensitive to the challenges that the author faced in the classroom. These students felt that “harsh” criticism could be viewed as discouraging for students that were facing challenges in the classroom. This feedback does highlight one drawback of the blinded process, which is the inability to know the learner’s characteristics in the blinded process and thereby preventing reviewers from providing feedback that is sensitive to learner’s characteristics. However, research has shown that an un-blinded process can lead to biased feedback based on stereotypes and inter-person dynamics (Cohen, Steele & Ross, 1999; Lu & Bol, 2007). Perhaps one way to keep the review process blinded and help students in providing supportive feedback is to make students explicitly aware of the criteria of good feedback along with exemplars of students’ reviews that meet the criteria.

For example, the data collected from the present round of implementation of the 4R Activity can be used to as exemplars for best practices in providing reviews. This information can then be provided to students along with many more opportunities to provide unbiased but constructive feedback.

b. The reflection activity

A support offered to students in building their competency in reviewing was the reflection activity, which was the last stage in the 4R Activity. This was done because unlike scientists, engineers and other professionals, students are not trained in the process of providing reviews and therefore need to be supported in this practice. In the reflection activity students re-read the review they provided to their peer in light of the scientist's review that the peer had received. Additionally students also read feedback that the scientist provided to each reviewer on how to improve his/her reviewing skills in the future. Students then reflected on aspects that they had learned from the exercise that they would then keep in mind in future reviewing activities.

When reporting on how the reflection helped students in building their peer reviews, students cited the scientist's feedback on reviewing skills, and having access to how the scientist provided the review. Findings pertaining to the improvement in elaboration and specificity in reviews over time and an increased uptake of characteristics from the scientists review (see Chapter 3, this dissertation) were previously recorded. In addition to these findings, students accounts captured by this study provide information on how students described the reflection activity as aiding in building their reviewing competencies. Student accounts provide information to support the argument made by many researchers that the access of expert's strategies (i.e., the scientist's review in this case) could significantly aid learners in learning that strategy (cf. Collins & Brown, 1988; Williams, 1993).

2. The 4R Activity and its relevance to students lives

This category encapsulates three themes that describe how students describe the applicability of the learning in the 4R Activity to their lives. The three aspects that the category encapsulates are what students find meaningful in the 4R Activity and how that relates to their present or future aspirations, the connections students draw between the practices in the 4R Activity and the practices of scientists, engineers and other professionals. Lastly, this category describes how students view the 4R Activity as a *tool* that can serve the purpose of providing a *link* between academic disciplines (e.g., science, mathematics, social studies) and their non-academic interests (e.g., music, athletics, dance). This category is in support of the argument put forth by Bell and colleagues (Bell et al., 2013) for the need to identify students' interests across social settings in order to able to support and promote development of expertise and interests of students. This development of expertise can then serve in fostering both academic and personal success.

a. Connections between the 4R Activity and the students' lives

Students' accounts on their learning from participation in the 4R Activity were closely tied to the importance of that learning in their present or future lives. These findings are important because it provide examples of how students use the agency in the science classroom to help in developing their own identity (Basu, Barton, Clairmont & Locke, 2011). For example, Student 1 considered *learning to provide feedback* as the most important aspect of participating in the 4R Activity. Moreover, Student 1 linked the usefulness *of learning to provide feedback* to her aspired career of wanted to be a teacher or a professor. Another student referred to learning *things that scientists do* and elaborated on wanting to be a scientist. In other words, these examples show that students used the 4R Activity to develop expertise in an area that they considered meaningful to their future lives as professionals.

Another example of students using the 4R Activity to develop expertise in an area that they consider important to their lives is the learning reported by the students classified as ELL. Both of the ELL students interviewed reported that their most important learning from the 4R Activity was the acquisition of scientific vocabulary, which provided them the agency to participate in future science classrooms. Thus the 4R Activity helped the ELL students in scientific language acquisition, which was meaningful to them because it allowed their learning to become more accessible to others in class. They reported that the acquisition of the scientific terminology would allow them to participate in classroom discussions more fully. These accounts reflect the claims made by scholars (e.g., Abedi, 2010; Meskill, 2010) who have discussed the dual learning challenge faced by ELL students of learning disciplinary content as well as English language simultaneously. The accounts of ELL students suggest that feedback on inclusion of scientific terminology as well as suggestions on how to use the scientific terminology in their written product might have aided in their learning of both the disciplinary content as well as the vocabulary in English language to express their disciplinary learning in class.

The students' accounts of learning as reported in the study provide a perspective to the research community on how students draw connections between their learning and the usefulness of that learning to their lives. The findings indicate that students desire learning environments that provide them with learning that they value as meaningful to their lives. This ties back to the idea expressed by Dewey (1902, 1938) of the need to provide students with opportunities to tie learning to their interests and what they consider as meaningful to their life. Additionally, it also ties back the vision of science education discussed in the *Framework* (NRC, 2012), which emphasizes the importance of making science relevant to the lives of the students. Furthermore and as outlined in research (e.g., Warren et al., 2001; Lee, 2006; Nasir et al., 2006)), identifying and effectively using students' experiences, interests and social knowledge in the science classroom can serve as a vehicle to promote learning. This study proposes that future research should continue to focus on identifying students' interests and experiences and effectively using it to promote learning in science. Additionally, students' accounts should be used to support students in drawing connections between the learning in the science classroom and its potential relevance to their lives.

b. Students connections between the 4R Activity and the work of professionals

The data from student interviews provided insight into students' perspectives on how they viewed the 4R Activity in relation to the practices of scientists and engineers. It was important to see whether students saw connections between the 4R Activity and the activities they perceived that scientists and professionals engaged in, such as engaging in receiving review and undertaking revision, providing review, and reflecting on feedback. As discussed previously in this paper, this was important because it provided insight into the ideas the students have about the work of scientists and also how they see the 4R Activity functioning in supporting them in doing similar work. With the NGSS's (NGSS

Lead States, 2013) emphasis on designing learning environments that reflect the practices of the real world scientists and engineers this information serves two purposes. First it helps in understanding the types of connections that students draw between the work of scientists and engineers and the learning environment (i.e., 4R Activity). Second, it provides us information on students' ideas on the work of scientists and engineers. This information can then be used in designing learning environments that support students in establishing connections between the activities undertaken by scientists and engineers and the activities that students engage in within the science classroom.

Interviews with students shed light on students' knowledge of the work of scientists and engineers through association with or direct observation of the work of these professionals, through engaging in multimedia content (e.g., internet, TV shows) and through the activities used in the science classroom (e.g., introductory video of the 4R Activity). These accounts showed that students are informed about the work of the science and engineering disciplines through a variety of sources, which equip them to participate in discussions about their learning in the science classroom. For example, students' accounts revealed that they considered knowledge building by scientists and engineers as a collective effort by the community, much the same way that students in the activity constructed NGSK products (Songer et al., in preparation) through receiving reviews and revising their constructed products. Their accounts also revealed that they considered engaging in peer review as not only providing unbiased feedback on their own written knowledge product but also serving as a rich source of learning about the work of others.

The connections that students drew between the 4R Activity and the work of scientists and engineers further suggests the nuanced understanding that students have about the work of professional scientists and engineers and how the practices in the science classroom are similar or different from the work of professionals. Additionally, the differences that students highlighted between the 4R Activity and the work of scientists and engineers (e.g., all students provided one prompt to respond to in the 4R Activity) revealed how students who looked at the activity from the lens of specificity could highlight important differences in how the 4R Activity differed from the way professionals engaged in such activities. These perspectives provide information on how to introduce the 4R Activity in successive iterations. For example, the 4R Activity could be introduced to the students through discussions that highlights the similarities and differences between the 4R Activity and the activities undertaken by scientists, engineers and other professionals. Furthermore, this study showcases students' accounts and how they interpret classroom activities in light of the work of scientists and engineers. These accounts can serve as a basis for further research as researchers continue to design activities that emphasizes learning in science classrooms to reflect how science is practiced by professionals (NRC, 2012; NGSS Lead States, 2013)

In addition to drawing connections between the 4R Activity and the work of scientists and engineers, students also described the applicability of the 4R Activity by other professionals. For example, student perspectives highlighted that they considered a number of other professionals (e.g., musicians, athletes) engaging in activities such as receiving review and undertaking revision, providing review, and reflecting on feedback. These findings could be interpreted through Brown, Collins and Duguid's (1989) claim that as learners begin to participate in authentic activities, they not only build a rich understanding of the tool within the context of use but also of the applicability of the tool in broader contexts. Thus as students engaged with the 4R Activity they began to see the connections of the activity to not only the work of scientists and engineers but also to the work of other varied professionals. Additionally, the drawing of connections to both science and non-science professionals could also be seen as fostering one of the goals of the *Framework* (NRC, 2012), which is to provide opportunities for students to develop skills that will enable them to be successful in a career (science or non-science) of their choice (NRC, 2012). This study highlighted how by initiating students to draw connections between the learning environment and other non-science professions allowed students to see the relevance of science activities undertaken in the classroom in non-science careers.

c. The use of the 4R Activity as an inter-disciplinary tool

Students' accounts provided information that students considered the 4R Activity helpful in their learning in science and a majority of students expressed a desire to see the use of the 4R Activity not only in science classrooms but in also other classes. They described scenarios of how the 4R Activity could be used in science classroom, in other subject areas (e.g., social studies, English Language Arts) as well their interests (playing video games, playing a musical instrument) that they engaged in outside school. Students expressed the presence of highly compartmentalized learning in different subject-areas and their desire to have a common link between different academic subjects and their life outside school. As described by one student, the 4R Activity could function as a *tool* to provide coherency to the lives of students in and out of school. Based on the students' accounts of a need to de-compartmentalize the learning associated with different subject-areas, this study supports the argument put forth by Stevens and colleagues (Stevens, Wineburg, Herrenkohl & Bell, 2005) on the need for an inter-disciplinary curriculum. Stevens and colleagues (2005) discuss the compartmentalization of subjects in the school day with little effort put into discussing the connections across different disciplines with students. They (Stevens et al., 2005) further highlight that the lack of attention paid into establishing inter-disciplinary connections leaves the important task of drawing these connections on the students. They argue for a new shift in school learning from “a fragmented collection of domain-specific accounts of school subjects to a comparative and unified one” (Stevens et al., 2013; p.136).

In an effort to promote such inter-disciplinary learning this study argues that learning environments like the 4R Activity can serve as an *inter-disciplinary tool* that can be adapted to be discipline specific while continuing to have some features that are discipline general. As an inter-disciplinary tool the 4R Activity would mediate the co-construction of knowledge while the process adopted in different stages of the 4R Activity (e.g., review, revise, reflect) could in turn be internalized by the learners and used in similar future activities (cf. Vygotsky, 1978; 1981). The use of the 4R Activity in different disciplines also could function as an opportunity to apprentice students in the discipline-specific norms of receiving and providing critique (cf. Lave & Wenger, 1991). Through the use of such an inter-disciplinary tool connections across disciplines as well as a comparative understanding of the disciplines can be facilitated. For example, the adaptation of the 4R Activity to English Language Arts would still have the four stages of the activity, but the prompts in each stage could be customized to be discipline specific.

An inter-disciplinary tool could also serve the purpose of providing connections between different disciplines for students much the same way that the standards documents provides connections for curriculum developers and educators between the standards in NGSS (NGSS Lead States, 2013) and the Common Core Standards (2010) in Mathematics and English Language Arts.

In summary, this study highlighted the affordances of using an activity like the 4R Activity to foster learning of NGSK. The student accounts showcased how an activity like the 4R Activity can be used to illuminate to students the collaborative nature of knowledge building in the scientific discipline. It also highlighted how students perceive learning environments designed to reflect the work of professional scientists and engineers. Future interventions designed to foster NGSK learning should perhaps include a component in their design of providing opportunities for students to reflect on the similarities and differences between the learning environment and the practices of scientists and engineers. It is through these classroom level discussions that educators can elicit student experiences and use them productively in further instruction. Additionally the discussions would allow students to reflect and learn more deeply about the work of scientists and engineers, which can further strengthen the connections students see between the designed activity and the work of professionals. Furthermore, these discussions can also be used as an avenue for students to connect their learning in science classrooms to the learning across different disciplines and to their everyday life.

References

Abedi, J. (2010). Research and recommendations for formative assessment with english language learners. In Andrade, H., & Cizek, G. J. (Eds.). (2010). *Handbook of formative assessment*. Routledge.

Basu, J., Barton, A.C.,Clairmont, N., & Locke, D. (2011).Youth Voices: Challenging the outcomes of science education. In S.J. Basu, A.C. Barton, and E. Tan (Eds.), *Democratic science teaching: Building the expertise to empower low-income minority youth in science* (pp. 21-40). Rotterdam, The Netherlands: Sense Publishers.

Bell, P., Bricker, L., Reeve, S., Zimmerman, H. T., & Tzou, C. (2013). Discovering and supporting successful learning pathways of youth in and out of school: Accounting for the development of everyday expertise across settings. In *LOST Opportunities* (pp. 119-140). Springer Netherlands.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (1999). *How people learn: Brain, mind, experience, and school*. National Academy Press.

Bricker, L. A. (2008). *A sociocultural historical examination of youth argumentation across the settings of their lives: Implications for science education*. ProQuest.

Cabaroglu N., Basaran S. & Roberts J. (2010). A comparison between the occurrence of pauses, repetitions and recasts under conditions of face-to-face and computer-mediated communication: a preliminary study. *Turkish Online Journal of Educational Technology, 9*(2), 4–23.

Cook-Sather, A. (2002). Authorizing students' perspectives: Toward trust, dialogue, and change in education. *Educational researcher*, *31*(4), 3-14.

Clark, C. (1995). *Flights of fancy, leaps of faith: Children' s myths in contemporary america*. University of Chicago Press

Collins, A., & Brown, J. S. (1988). *The computer as a tool for learning through reflection* (pp. 1-18). Springer US.

Cohen, G. L., Steele, C. M., & Ross, L. D. (1999). The mentor's dilemma: Providing critical feedback across the racial divide. *Personality and Social Psychology Bulletin*, *25*(10), 1302-1318.

Dewey, J. (1902). *The child and the curriculum.* Chicago: The University of Chicago press.

Dewey, J. (1938). *Experience and education.* New York: Macmillan.

Deakin, H., & Wakefield, K. (2013). Skype interviewing: Reflections of two PhD researchers. *Qualitative Research*, 0 (0), 1-14. doi: 10.1177/4687941/3488/26

Denscombe M. (2003). *The Good Research Guide*. Maidenhead: Open University Press.

Elliot, S.N., Kettler, R.J., Beddow, P.A., Kurz, A. (2010). Research and strategies for adapting formative assessments for students with special needs. In Andrade, H., & Cizek, G. J. (Eds.). (2010). *Handbook of formative assessment*. Routledge.

Elen, J., & Lowyck, J. (1998). Students' views on the efficiency of instruction: An exploratory survey of the instructional metacognitive knowledge of university freshmen. *Higher Education, 36*, 231–252.

Elen, J., & Lowyck, J. (1999). Metacognitive instructional knowledge: Cognitive mediation and instructional design. *Journal of Structural Learning and Intelligent Systems, 13*(3–4), 145–169.

Emerson, Fretz & Shaw (2011). *Writing ethnographic fieldnotes (2nd edition).* Chicago: University of Chicago Press.

Entwistle, N. J. (1991). Approaches to learning and perceptions of the learning environment. *Higher Education, 22*(2), 201–204.

Finders, M. (1997). *Just girls: Hidden literacies and life in junior high.* New York: Teachers College Press.

Guion, L. A., Diehl, D. C., & McDonald, D. (2011). Triangulation: Establishing the validity of qualitative studies.

Hanna P (2012) Using Internet technologies (such as Skype) as a research medium: a research note. *Qualitative Research, 12*(2), 239–242.

Hulleman, C. S., & Harackiewicz, J. M. (2009). Promoting interest and performance in high school science classes. *Science*, *326*(5958), 1410-1412.

Johnston, P. & Nicholls, J. (1995). Voices we want to hear and voices we don't. *Theory into Practice,* 43 (2), 94-100.

Harackiewicz, J. M. (Eds.)(2000). *Intrinsic and extrinsic motivation: The search for optimal motivation and performance*. Academic Press.

Lepper, M.R. & Henderlong, J. (2000). Turning" play" into" work" and" work" into" play": 25 years of research on intrinsic versus extrinsic motivation. In C. Sansone & Harackiewicz, J. M. (Eds.). (2000). *Intrinsic and extrinsic motivation: The search for optimal motivation and performance*. Academic Press: San Diego.

Lincoln, Y. (1995). In search of students' voices. *Theory into Practice, 43*(2), 88-93.

Lu, R., & Bol, L. (2007). A comparison of anonymous versus identifiable e-peer review on college student writing performance and the extent of critical feedback. *Journal of Interactive Online Learning*, *6*(2), 100-115.

Liu, N. F., & Carless, D. (2006). Peer feedback: the learning element of peer assessment. *Teaching in Higher education*, *11*(3), 279-290.

Lundstrom, K., & Baker, W. (2009). To give is better than to receive: The benefits of peer review to the reviewer's own writing. *Journal of Second Language Writing*, *18*(1), 30-43.

Lu, R., & Bol, L. (2007). A comparison of anonymous versus identifiable e-peer review on college student writing performance and the extent of critical feedback. *Journal of Interactive Online Learning*, *6*(2), 100-115.

Lyons, T. (2006). Different countries, same science classes: Students’ experiences of school science in their own words. *International Journal of Science Education*, *28*(6), 591-613.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge university press.

Maxwell (2013). *Qualitative research design: An interactive approach.* Thousand Oaks, CA: Sage.

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

McDermott, R. P. (1996). The acquisition of a child by a learning disability. In S. Chaiklin & J. Lave (Eds.), Understanding practice: Perspectives on activity and context (pp. 269–305). Cambridge, UK: Cambridge University Press.

Meskill, C. (2010). Moment-by-moment formative assessment of second language development: ESOL professionals at work. In Andrade, H., & Cizek, G. J. (Eds.). (2010). *Handbook of formative assessment*. Routledge.

Nasir, N. I. S., Rosebery, A. S., Warren, B., & Lee, C. D. (2006). Learning as a cultural process: Achieving equity through diversity. *The Cambridge handbook of the learning sciences*, 489-504.

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common Core State Standards*. Washington, DC: Authors.

National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas.* Committee on Conceptual Framework for the New K-12 Science Education Standards. Board on Science Education. Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

National Research Council. (2014). *Developing Assessments for the Next Generation Science Standards.* Committee on Developing Assessments of Science Proficiency in K-12. Board on Testing and Assessment and Board on Science Education, J.W. Pellegrino, M.R. Wilson, J.A. Koenig, and A.S. Beatty, *Editors.* Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

NGSS Lead States (2013). *Next Generation Science Standards: For States, By States.* Washington, DC: The National Academies Press.

Purchase, H. C. (2000). Learning about interface design through peer assessment. *Assessment & Evaluation in Higher Education*, *25* (4), 341–52.

Resnick, L. B. (1987). The 1987 presidential address: Learning in school and out. *Educational researcher*, 16(9), 13-54.

Resnick, L. B. (1986). The development of mathematical intuition. In M.Perlmutter (Ed.) (1986). *Perspectives on intellectual development: The Minnesota Symposium on Child Development, 19*, 159-194.Hillsdale, NJ: Erlbaum.

Rodgers, C. R.. (2006). Attending to Student Voice: The Impact of Descriptive Feedback on Learning and Teaching. *Curriculum Inquiry*, *36*(2), 209–237. Retrieved from http://www.jstor.org.proxy.lib.umich.edu/stable/3698505

Sedgwick M and Spiers J (2009) The use of videoconferencing as a medium for the qualitative interview. *International Journal of Qualitative Methods, 8*(1), 1–11.

Songer, N.B., Kelcey, B., & Gotwals, A. (2009). How and when does complex reasoning occur? empirically driven development of a learning progression focused on complex reasoning About Biodiversity. *Journal of Research in Science Teaching, 46*(6), 610 - 633.

Songer,N.B., & Gotwals, A.W. (2012). Guiding explanation construction by children at the entry points of learning progressions. *Journal of Research in Science Teaching, 49(2),* 141–165.

Songer, N.B., Dewey, T., Peters, V., Reicher, M., Kwok, A., Zaidi, S., Tupper, B., Hammond, G., Myers, P. (2014) *Climate Change and Impacts on Ecosystems: A Middle School Curricular Unit.* Ann Arbor, MI: The University of Michigan.

Stevens, R., Wineburg, S., Herrenkohl, L. R., & Bell, P. (2005). Comparative understanding of school subjects: Past, present, and future. *Review of Educational Research*, *75*(2), 125-157.

Teo, A. K. (2006). Social-interactive writing for english language learners. *The CATESOL Journal, 18*, 160–178.

Vygotsky L. 1981. The instrumental method in psychology. In *The Concept of Activity in Soviet Psychology,* ed. J Wertsch, pp. 3–35. Armonk, NY: Sharpe.

Vygotsky L. 1978. *Mind in Society: The Development of Higher Psychological Processes,* ed. M Cole, V John-Steiner, S Scribner, E Souberman. Cambridge, MA: Harvard Univ. Press.

Weiss, R. (1994). *Learning from strangers: The art and method of qualitative interview studies.* New York: Free Press.

Williams, S. M. (1993). Putting case based learning into context: examples from legal, business, and medical education, Journal of the Learning Sciences, *2*, 367-427.

Weinmann, T., Thomas, S., Brilmayer, S., Heinrich, S., & Radon, K. (2012). Testing Skype as an interview method in epidemiologic research: response and feasibility. *International journal of public health*, *57*(6), 959-961.

Warren, B., Ballenger, C., Ogonowski, M., Rosebery, A. S., & Hudicourt-Barnes, J. (2001). Rethinking diversity in learning science: The logic of everyday sense-making. *Journal of research in science teaching*, *38*(5), 529-552.

Chapter 5

Conclusion

This dissertation presented an assessment system called the *4R Activity* that was grounded in the vision of classroom assessments for NGSS (NRC, 2014). The assessment report (NRC, 2014) emphasizes that classroom assessments designed to support integrated knowledge should provide students with: (a) multiple opportunities to deepen NGSK, (b) guidance to advance their science learning and (c) opportunities to reflect on their performance (NRC, 2014). Three studies on the 4R Activity were presented in this dissertation. The three studies used varied research methods to address key questions of efficacy of the 4R Activity in supporting NGSK learning, progressions in learning observed during the 4R Activity and the experiences of students who participated in this novel assessment system. The contributions made by the three studies are discussed below.

1. Customized feedback and revision provided opportunities to support students in developing proficiency on Next Generation Science Knowledge (NGSK)

This dissertation provided evidence for the efficacy of the 4R Activity in supporting student learning of NGSK (see Chapter 2, this dissertation). The study's primary contribution was to shed light on how the 4R Activity can be used to accelerate learning of *all* students while also serving the needs of underrepresented students. This was done by statistically comparing the achievement gains of students using the 4R

Activity with those students who used a conventional assessment activity (1R Activity). The comparison of achievement gains was made at three time points (pre, mid and post) using a fixed-effect regression model. This study demonstrated that group assignment had the most significant effect on achievement gains. In the 4R Group, *all* students outperformed (mean=2.29 points) the students in the 1R Group with the difference in achievement gains between the two groups being largest (~2.6 times) for more difficult NGSK items. Moreover the results indicated that in addition to the positive main effect of the 4R Activity on achievement gains in the mid-to-post period, there were statistically significant heterogeneous effects of 4R Activity on females, African Americans and other non-white racial/ethnic race groups. This resulted in traditionally underrepresented sub-groups in STEM (NRC, 2011) having an additional positive increase in scores (mean increase=0.80 points) in the mid-to-post period. Additionally, the 4R Activity assisted 11% of students who began with low levels on NGSK (a score of 20% or lower) to reach proficiency levels (a score of 80% or higher) by the end of the curriculum. This was in contrast to the students in the 1R Group where less than 1% of students with low NGSK were able to reach proficiency levels by the end of the curriculum.

The results lend support to the use of customized feedback along with revision opportunities to promote learning of NGSK. Customized feedback was closely tied to a student's NGSK product and provided them with key areas that needed to be addressed to strengthen the NGSK product. Furthermore, the revision activity provided students with an opportunity to engage with the feedback more deeply than they would have if they did not have to revise the knowledge product.

Given that all students received feedback and were encouraged to revise, students' perceptions on "who receives feedback" and "who is asked to revise" might have also changed over time (see Chapter 3, this dissertation). The emphasis on revision might have helped students see that revision did not necessarily imply that the original knowledge product was unsatisfactory (see Chapter 3, this dissertation) but that further improvements could be made to a knowledge product to improve its quality. The lack of score assignment in the received feedback might have lead to an increased motivation of students to engage with the feedback and revise their knowledge products. The increased motivation of students to engage with feedback in the absence of grades has also previously been documented in research (e.g., Black & Wiliam, 2004). Additionally, as suggested by the data on incorporation of feedback (see Chapter 3, this dissertation) perhaps more students became open to accepting feedback over time and engaged with it to revise their knowledge products.

2. Formative assessments is a productive vehicle to promote discipline specific activities

The *Framework for K-12 Science Education* (NRC, 2012) advocates for development of integrated knowledge (e.g., NGSK) in students through learning environments that illuminate to students the work of scientists and engineers (NRC, 2012). Responding to this need, Chapter 3 in this dissertation discussed how students not only constructed NGSK products but also engaged in *critique*, an activity central to the work of scientists and engineers. The 4R Activity supported students in building NGSK while also building their competency in receiving feedback from peers and an expert and providing feedback to their peers through a double-blinded review process. Additionally, students were supported in building competency in reviewing through a reflection activity. This chapter documented the emerging capabilities of students across the three 4R Activities in the areas of: (a) revision of NGSK products, (b) providing feedback, and (c) self-reflection on reviewing skills. The chapter discussed how with each successive 4R Activity more students were able to reach proficiency levels on their revised NGSK products by the end of the activity. Additionally, students' reviews became specific and elaborated over time and reflected the characteristics observed in the scientist's reviews. This study also showed how students' reflections on their own peer reviewing skills became detailed with most students including a self-assessment of their reviewing skills and elaborating on the aspects that they would like to improve upon when providing reviews in the future. This study's major contribution was in showcasing that students can be supported in building proficiency in NGSK through the process of receiving and providing critique, which is an essential element of science learning (see NRC, 2012). Additionally, the study paves way for future work in the area of examining how increased competency in engaging in critique can contribute to increased proficiency on the developed NGSK product.

3. Students' experiences and interests about classroom formative assessments provide important information about the value of the assessment

Chapter 4 in this dissertation contributed to research on the use of classroom assessments to elicit students' experiences to inform instructional goals and promote student learning. Eliciting of students' experiences through classroom assessments has traditionally not been a focus in assessment research (NRC, 2014). However the assessment report (NRC, 2014) advocates for the use of assessments to support student learning (NRC, 2014). Responding to this need, Chapter 4 presented the accounts of twenty-two students. It highlighted their perspectives on the learning they found valuable while engaging in the 4R Activity. A majority of students interviewed expressed that learning the process of "receiving and/or providing feedback" was the most valuable outcome of the activity. The study also documented the connections students saw among the different aspects of the 4R Activity and their interests, aspired careers and the work of scientists and engineers. The findings showcase how students' leveraged their interests (e.g., playing basketball) to draw connections to the activities (i.e., 4R Activity) they do in the science classroom. Additionally, the findings also highlight students' perceptions about the work of scientists and engineers and how students saw the different stages within the 4R Activity (Respond, Review, Revise Reflect) applicable to the work of professional scientists and engineers. These findings provide insight into how classroom assessments can be used to leverage students' experiences, which can then be used effectively to further instructional goals and promote learning.

4. There is value in resources that support students' with agency in their own learning

Agency can be defined as providing individuals the ability to make choices and act on those choices (Martin, 2004). Research on learning environments that provide students with agency in their learning have shown positive learning outcomes (see Corbett, Koedinger & Anderson, 1997; Biswas, Leelawong, Schwartz & Vye, 2005).

Additionally, in the area of formative assessments, providing agency to students is often cited as an important tenet of effective formative assessments (Black & Wiliam, 1998; Wiliam &Thompson, 2007). For example, in Wiliam and Thompson's (2007) framework on strategies for successful formative assessments, two out of the five strategies relate to promoting student agency. These strategies are: (a) activating students as instructional resources for one another, and (b) activating students as the owners of their own learning (Wiliam & Thompson, 2007).

The framework of Wiliam and Thompson (2007) was used to guide the design of the 4R Activity (see Chapter 1, this dissertation). The 4R Activity used the above outlined strategies in the formative assessment design and thus provided students with a sense of agency in their own learning. In the 4R Activity students were provided agency through peer review, the revision and reflection processes. Peer review made student feedback as a central piece in the construction of the NGSK products. It allowed students to productively contribute to knowledge building of their peers while enabling students to provide the type of feedback to their peers that they deemed most important. Additionally the process of revision provided agency to the students by allowing them to make revisions on aspects that the student deemed as necessary. Furthermore, the process of reflection gave an opportunity to students to reflect on their own learning of reviewing. The reflection activity guided students in reflecting on aspects of his/her reviewing skills that they considered important to build upon in future reviewing activities. Through this the reflection activity contributed to students' sense of agency in their learning.

Student agency was also explored in this dissertation through the documentation of students' accounts (see Chapter 4, this dissertation). Interviews with students can be seen as promoting student agency because the interview was an opportunity to hear students' experiences and the relation of the experiences to the different aspects of the 4R Activity. The semi-structured interviews allowed students to reflect on the learning from engaging in the 4R Activity, draw connections between their interests and the different stages of the 4R Activity and also reflect on the connections they saw between the different stages of the 4R Activity and the work of scientists and engineers. Additionally, suggestions for improvements to the 4R Activity were also collected from students. This gathered feedback along with the findings from the interview data together used students' accounts to impact future iterations of the 4R Activity.

5. Classroom formative assessments can promote opportunities to learn for all students

Providing students with a fair opportunity to learn is a pressing challenge in U.S. education (NRC, 2012). While the challenge at the national level concerns with the inequities among schools, districts and states with respect to teacher preparedness, instructional supports, and material resources (NRC, 2012), this dissertation contributed in addressing the challenge at the classroom level. In this dissertation a fair opportunity to learn for all students was promoted through a formative assessment, the 4R Activity. Furthermore, this dissertation demonstrated that positive learning outcomes can be achieved for all students when formative assessment systems provide fair opportunities to learn to all students. In the 4R Activity providing opportunity to learn for all students was promoted through the design of the activity. The findings suggest that the design of the 4R Activity was a contributing factor in the positive learning outcomes demonstrated by students. The 4R Activity strived to provide opportunity to learn for all students through customized feedback, and peer feedback. Customized feedback acknowledged that students were at different proficiency levels on NGSK and provided feedback that could help students achieve higher proficiency on the target NGSK. This was in contrast to the classroom discussion format, which is a common practice of providing feedback to students in science classrooms (as reported by all participating teachers). In classroom discussions the feedback is not tied to students' individual proficiency on the NGSK but is focused on the proficiency of the class as a unit. Additionally it is difficult to monitor if all students are sufficiently engaged with the feedback. Through customized feedback and revision opportunities, all students received feedback that was closely tied to their knowledge products and received an opportunity to engage with the feedback in the revision and resubmission of the NGSK product.

The second way in which a fair opportunity to learn for all students was promoted was anonymous peer feedback. Anonymous peer feedback placed students as an important instructional resource in the knowledge building process. Furthermore, providing feedback was not reserved for only high performing students who functioned as tutors for low performing students. In the 4R Activity, all students got an opportunity to review the work of others and provide peer feedback. Additionally, the peer feedback was anonymous (e.g., double-blinded), which helped in eliminating bias in feedback that might have occurred due to inter-person dynamics or gender/racial stereotypes. The blinded process also prevented students from providing constructive feedback to only their friends (see Chapter 4, this dissertation). Classroom observations and students' accounts also revealed that students appreciated having an opportunity to provide feedback to their peers and indicated that they learned both NGSK and peer reviewing skills through the process (see Chapter 4, this dissertation). As suggested in research (DeGuerro & Villamil, 2000; Teo, 2006; Lundstrom & Baker, 2009), providing feedback to peers has shown to be beneficial to the reviewer as it provides the reviewer with access to the work of his/her peers and allows the reviewer to constructively engage with the peer's work. Thus, in the 4R Activity, engaging in the process of anonymous peer review provided all students with an opportunity to learn from the work of their peers. In summary, the design of the 4R Activity promoted a fair opportunity to learn for all students who engaged in the activity.

Limitations

The studies discussed in this dissertation all stemmed from the use of the 4R Activity and collectively presented learning outcomes and students' perceptions of learning from participating in a formative assessment. However the studies have some limitations that stem from the sampling procedures, the amount of data collected, and the type of NGSK assessed.

Sampling Procedures

The sampling procedure adopted for selection of participants for the dissertation was not based on random sampling procedures but based on inclusion of schools that had volunteered to participate in the study. From the list of school sites, measures to balance the control (1R Group) and intervention group (4R Group) were based on total number of students, grade- span (grade 6-8), number of English Language Learners and number of students with IEPs. This resulted in four school sites being placed in the intervention group while there was only one school site in the control group. Having only one school in the control group limits the power of the results pertaining to the learning gains reported in Chapter 2. In addition to the measures taken to balance the sample, balancing the sample with respect to the number of participating schools would have provided more power to the statistic models run.

Furthermore, a fixed effects model was run to compare the achievement gains of the control and intervention group students. One disadvantage of the fixed effects approach is that the results are conditional on the data used to estimate them and cannot be generalized to other samples not included in the study. Thus the results reported in Chapter 2 are not generalizable to a larger population.

Moreover the small sample of students classified as English Language Learners and students with IEPs restricts the interpretation of the learning gains for these populations. A larger sub-sample of these two sub-populations would allow researchers to better explore the effects of the 4R Activity in these sub-samples.

Data Collected

Another limitation of the 4R Activity implementation was that there were only three time points of implementation of the activity. This restricted the analysis in some ways. For example in Chapter 2, the small number of time points did not allow for the robust examination of the effects of participating in a prior 4R Activity on future 4R Activity participation. Also in Chapter 3 an analysis of uptake of feedback on reviewing skills and utilization of reviewers' reflections on the reviewing skills was not carried out.

Moreover, the study used a posttest administered right after curriculum completion to provided evidence of short-term learning gains. Perhaps future research can also look at long-term learning gains of the NGSK by administering a second posttest at a future time, (e.g., several months after curriculum completion). This will provide the much needed insight into the long-term learning effects of participating in the 4R Activity.

With respect to the data collected for student interviews (Chapter 4) the data was collected at only one time point soon after the completion of the curriculum. Multiple time points, such as after the first 4R Activity, at the end of the curriculum (post interview), and a few weeks after the end of the curriculum (post-post interview) would have allowed for an examination of progressions in students' thinking. Furthermore, a follow up post-post interview would have provided a more nuanced understanding of the types of modifications that some students make to the different aspects of the 4R Activity if/when they apply it to activities outside the science classroom. This would have allowed for a more complete picture of if/how students apply their experiences from the science classroom to their everyday lives.

Type of NGSK Assessed

This study looked at construction of only one type of NGSK that used the science practice of constructing explanations. The findings reported in this study by no means claim the use of the 4R Activity as the best method to assess and support all types of NGSK. The results do not provide evidence of the use of the 4R Activity to assess and support other types of NGSK that might include different disciplinary ideas, practices and crosscutting concepts. However, some suggestions are made in forthcoming sections of this chapter on the types of NGSK that might be suitable for use with the 4R Activity.

Further research using different types of NGSK will provide more evidence of the types of NGSK that are best assessed and supported by the 4R Activity.

Future Work

Future work with respect to the 4R activity stem from the findings and limitations discussed in this dissertation as well as from the suggestions received from the five teachers and over 180 students that used the 4R Activity.

Improvements to successive iterations of the 4R Activity

There were a number of design improvements that emerged directly from the classroom observations conducted, the limitations of the studies, the challenges reported by students and teachers, and the improvements suggested by the students. The design improvements presented below are further categorized into three categories: (i) addition of features to the 4R Activity system; (ii) addition of features to aid in scalability of the 4R Activity; and (iii) inclusion of activities in conjunction to the 4R Activity.

1. Addition of Usability Features to the 4R Activity

A number of features that could enhance the usability of the 4R Activity are presented below.

a. Feature for asking clarification questions to reviewers. When conducting interviews with students, some students reported that they did not prefer the blinded review process because not knowing the identity of the reviewer prevented them from asking clarifying questions. However the double-blinded process worked well for many other students and has numerous advantages previously reported in research (see Chapter 4, this dissertation). In order to address this challenge while continuing to maintain the double-blinded process, an optional feature of asking clarifying questions to the reviewer could be added. This feature would help students to directly communicate with the reviewer and seek clarification on unclear aspects in the review. Similarly this feature would also allow students to asking clarifying questions to the scientist. Through this feature students could receive further help with the revision process.

b. Text-to-speech capability for reading feedback. During the implementation of the 4R Activity some teachers assisted English Language Learners with reading the reviews to support them in the revision process. This sometimes took away valuable time of the teacher which could have been used to help students with other aspects such as how to effectively incorporate the feedback in their revised NGSK product. In response to this the teachers as well as the students (including a students classified as an ELL) suggested that an optional text-to-audio feature could help in making the 4R Activity more equitable for students who might face some challenges in reading written feedback. Such a feature could free teacher time from helping students read the feedback to allowing them to help students with specific aspects of implementing the feedback.

c. Feature to ask for teacher assistance. Drawing from suggestions given by students with IEPs, it was suggested that they would prefer a button where they could request help from the teacher. This would allow teachers to instantly know who had requested help and provide assistance accordingly. Furthermore the feature could provide a drop down menu of areas of help to choose from such as, if help was needed with understanding activity instructions, or understanding feedback, or to report a technology glitch.

Additionally the data on what students seek help on could be valuable in further improving the 4R Activity.

d. Feature to keep a running list of "aspects to improve upon in providing peer review". Many students reported that one of the biggest takeaways from the 4R Activity was how to provide peer reviews. As a suggestion to further build peer-reviewing skills some students reported that it would be beneficial to have a running list of "things to improve upon when providing peer review". This list would be populated with students' prior reflection activities and could easily be made available as a resource in the review stage of the 4R Activity. Additionally students also suggested that they would have liked the opportunity to give feedback to the reviewer on the helpfulness of his/her review. This feedback could also be added to the list of improvements n peer-reviewing skills. The author could provide feedback by choosing from a list of pre-defined areas. For example, "suggested examples of edits would be helpful", or "specifying the location of the error would be helpful". The author could then select all areas that they thought applied to the reviewer.

2. Features for Promoting Scalability of the 4R Activity

Scaling the 4R Activity to accommodate many more participating schools would require rethinking some of the present functionality of the 4R Activity, such as the role of the scientist in providing feedback to all students. To make the 4R Activity scalable, the role of the scientist could be performed by an automated scoring system.

Automated scoring is defined as the use of non-human subjects (i.e., a computer system) to evaluate and score student products (Shermis & Burstein, 2003). There is a growing body of research on using automated scoring for essays in English language Arts and Social Studies. Many automated scoring systems exist that are presently being used by testing agencies. Some of them are, Project Essay Grader ™ (PEG) (Page, 2003) used by College Board, Intelligent Essay Assessor ™ (IEA) produced by the Pearson Knowledge Technologies (2010), e-rater® developed by Educational Testing Service (ETS) (Burstein, 2003) and Bayesian Essay Test Scoring™ (BETSY) developed by Rudnger and Liang (2002) for the Graduate Management Admission Council (GMAC). While the automated scoring systems mentioned above differ in the specific methods employed to score responses, they all include training the model to quantify the features of the essay to predict the score that would have been given by a human scorer.

Using an automated scoring system in the 4R Activity could allow assignment of peer reviewers who were not randomly assigned but assigned based on the principle of *zone of proximal development* (Vygotsky, 1978) of the student. The *zone of proximal development (ZPD)* is defined as, "the distance between the actual developmental level as determined through independent problem solving and the level of potential development as determined through problem solving under adult guidance or in collaboration with more capable peers" (Vygotsky, 1978; p.86). Learning that takes place in the ZPD of a person in the presence of knowledgeable others can help the learner achieve success in the learning activity, in a way that they could not have done without the support. When using automated scoring system, first the original NGSK product of students (from the first stage of the 4R Activity) would be scored using the holistic rubric (Songer, Zaidi & Newstadt, in preparation). Based on the score of the NGSK product, reviewers could be assigned so that the reviewer's score on his/her own NGSK product is only one or two levels above the score of the NGSK product they are assigned to review.

The new role of the scientist could be taken up the computer system, which would use the score to assign feedback strategies from a corpus of pre-developed strategies for each NGSK product. The use of such a system is in line with the work of several other researchers (e.g. Kennedy & Wilson, 2007; Minstrell & VanZee, 2003). For example Kennedy and Wilson (2007) proposed the use of progress variables that categorized students based on their levels of understanding. The teachers then used the categorization to suggest next steps. Similarly, Minstrell (2003) had also proposed the DIAGNOSER system used in Physics that suggested activities and next steps to students after categorizing students based on the data collected from the formative assessment.

While replacing the role of the scientist by a computer could ensure efficiency and help in scalability, it might lead to reduced motivation of students engaging in the 4R Activity. As documented in students' accounts, many students' reported to have learned from the scientist's reviews on how to be a stronger reviewer and how to revise their NGSK product (see Chapter 4, this dissertation). Additionally, classroom observations and discussions with teachers also revealed that students were motivated to engage in the 4R Activity because they felt that a scientist at a prestigious university was reviewing their work and providing them with feedback.

Maintaining the role of the scientist to provide feedback would require establishing a community of professionals who participate in the 4R Activity. To aid in establishing such a community, organizations could incentivize employees to participate in the program. Additionally, professional development programs would need to be established to support and prepare participating professionals in providing feedback on student work and providing mentorship to students. While considerable efforts would be required to set up such a community, it would aid in promoting student learning and providing students with access to professionals in the science and engineering community. Establishing a community of scientists and engineers who partner with schools would allow students to not only receive feedback in the 4R Activity but the scientist-student partnership could also be used to help students better understand the work of STEM professionals. The scientist-student partnership would provide access to students who might otherwise not have access to observing the work of scientists or engineers. This might help in cultivating student interest in STEM fields and also provide a nuanced understanding of the work of professionals in STEM. Thus, if scientist-student partnership programs can be established, scalability of the 4R Activity could be achieved without compromising on aspects of student learning that might have resulted due to the involvement of a scientist in the 4R Activity.

3. Additional Activities to be done in conjunction with the 4R Activity

Three types of activities to be administered in conjunction with the 4R Activity are suggested. These activities are suggested to support students in providing reviews and incorporating revisions, provide opportunities for students to discuss the relevance of the 4R Activity in their life and in the science and engineering disciplines, and to act as a building block to deepen their learning of NGSK.

In addition to emphasizing the role of revision in scientific knowledge building there is a need to discuss the importance of the 4R Activity in the science and engineering fields as well as to help students with drawing connections between the 4R Activity and their interests. Some activities that could be introduced in conjunction with the 4R Activity are described below.

a. Providing opportunities to receive and provide peer review outside the 4R Activity. The suggestions from the participating students and teachers, as well as from the studies conducted on the 4R Activity, revealed that there was a need to include additional activities to support students with the peer review and revision aspects of the 4R Activity. As discussed in Chapter 3, students provided stronger peer reviews in classrooms where teachers used the peer reviewing process from the 4R Activity in other lessons in the curriculum (e.g., Ms. Asha and Ms. Lata). Thus more activities that used peer review would help students in developing competency in this process.

Furthermore, students need practice with understanding the norms of providing constructive feedback. As discussed in Chapter 4, some students had reported as not preferring the blinded process because they felt that they did not want to be “harsh” to students without knowledge of the prior challenges faced by that student in the science classroom. This sheds light on an important aspect that students need more opportunities to provide constructive feedback regardless of the person who is receiving the feedback. Explicitly establishing some classroom norms on how to give constructive feedback would be beneficial to students. For example, activities on providing feedback followed by a discussion on strategies to provide constructive feedback coupled with showing students’ exemplar feedback products will help students in developing competency in providing feedback to their peers. Furthermore, it will equip students to provide feedback in a constructive manner instead of providing unconstructive feedback or not choosing to provide feedback due to the fear of coming across as “harsh”.

With respect to incorporating feedback from the reviews two things were brought to light. First many students were hesitant in revising their NGSK product in the first 4R

Activity because they felt that revision implied that their original NGSK product was not satisfactory (see Chapter 3). Second, most students had no previous exposure to revising in science classrooms and did not how to effectively go about it (see Chapter 3). These aspects highlight that the role of revision during knowledge building in science needs to be emphasized and more activities that include revision need to be included. In further implementation cycles of the 4R Activity, it would be beneficial to provide teachers with some guidelines on how they could use the peer review activity more often in their classes.

b. Providing opportunities to discuss relevance of the 4R Activity. Students expressed the desire to have opportunities to discuss how they saw the 4R Activity in relation to their interests, everyday lives and the work of professionals (see Chapter 4). In the present implementation of the 4R Activity, an informative video on the 4R Activity was screened in classrooms prior to students doing the 4R Activity. In addition to explaining the 4R Activity, the video briefly covered aspects of the use of the different aspects of the 4R Activity by scientists, engineers and other professionals (e.g., fashion designers, editors). In future implementation of the 4R Activity a discussion following the video could serve in helping students discuss the connections between the different aspects of the 4R Activity to the work of professionals as well as to their lives. The discussion could have students brainstorm additional professionals that use the 4R Activity as well as discuss the particulars of how the different aspects of the 4R Activity would be used by professionals in their disciplines. The discussion would also include students brainstorming on different ways of using some or all aspects of the 4R activity in their activities of interest. Furthermore, discussions as the one described above should occur at multiple time points in the course of a school year to facilitate meaningful connection building between the science classroom and the lives and interests of the students. These discussions can further aid students in more deeply understanding the work of scientists and perhaps help more students in choosing STEM careers.

c. Extension activities to deepen NGSK. The vision of formative classroom assessments is to both assess and support learning. While much work has been done in the field on the "assessing" aspect of assessments, there is still work required in understanding how assessments can function as a vehicle to support learning (Andrade & Cizek, 2010). The 4R Activity provided students with an opportunity to iteratively build their NGSK products. While many students were able to reach proficiency levels on their revised NGSK (see Chapter 2) not all students were able to demonstrate proficiency on the assessed NGSK by the end of each 4R Activity. Thus activities to further support students in developing proficiency in NGSK should be provided. In an effort to further support the learning of NGSK, additional activities succeeding the 4R Activity should be designed. These activities will allow students to revisit the NGSK products constructed as well as provide further support in fostering their learning of the NGSK. As discussed in the *Framework* building NGSK requires multiple iterations and repeated exposure (NRC, 2012). To do this Pellegrino and colleagues (NRC, 2014) suggest engaging students in multiple scientific practices used in concert with one another to develop proficiency in NGSK. Building on this recommendation an activity to deepen students' NGSK could be employed after the 4R Activity that engages students in argumentation structured around the revised NGSK product. The *Framework* (NRC, 2012) outlines multiples ways to support middle school students in *engaging in argument from evidence* (Practice 7, NGSS Lead States, 2013). Of the many ways, the revised NGSK product from the 4R Activity serves well in supporting students in one way:

> *Respectfully provide and receive critique about one's explanations, procedures, models, and questions by citing relevant evidence and posing and responding to questions that elicit pertinent elaboration and detail.* (Appendix F NGSS Lead States, 2013; p. 13)

By the end of the 4R Activity each student has a revised NGSK product, which can then be used to engage students in argumentative discourse where their peers can pose questions and students can respond with further elaboration. By engaging in a classroom activity on a given explanation, students will be made aware of the variety of evidence and reasoning that their peers use to justify a claim. This can deepen the knowledge of NGSK and help students with challenging aspects of explanation construction, such as difficulty with using appropriate evidence (e.g., Sandoval & Millwood, 2005), and understanding what counts as reasoning to support the claim and evidence (e.g., Gotwals & Songer, 2009; Bell & Linn, 2000).

Closing Remarks

This dissertation presented an example of a formative assessment system that supported student learning of NGSK through the use of customized feedback and revision opportunities provided to students. Given that the assessment was online, it was possible to administer an anonymous review process and prevent potential bias in feedback. This customized and anonymous feedback process along with the opportunity to revise NGSK products was perhaps a significant contributor to the high achievement gains on NGSK demonstrated by students. While supporting students in developing NGSK, the 4R Activity also apprenticed students into the activities of review, reflection and revision.

Additionally, this dissertation made an important first step in documenting students' ideas and experiences to gain insight into productively leveraging students' experiences to inform student learning and design decisions of learning environments. Based on the findings from this dissertation some implications for research and practice are offered.

Implications for Science Educators

One of the most salient implications of this work for science educators is the need to foster and support "productive interactions" in the science classroom. "Productive interactions" can be defined as interactions between participants in which contributions (e.g., knowledge, experiences, interests, ideas) from all participants are legitimately valued and leveraged to inform learning and teaching (Cameratti-Baeza, in preparation). The author of this dissertation proposes two forms of productive interactions for science classrooms: (a) productive interactions between the teacher and students; and (b) productive interactions among students.

Productive interactions between the teacher and students can be facilitated when opportunities are provided to students in the class to draw on their expertise from outside the science classroom. In order to facilitate productive interactions between students and teacher, both in-service and pre-service teachers need to be supported in effectively leveraging students' interests, especially when they might seem unrelated to the topic of discussion. Professional development programs would need to be set up to support teachers in designing classroom instruction and assessments that promote productive interactions. Additionally emphasis also needs to be placed by teachers on leveraging students' interests and community practices (see, Warren et al., 2001; Nasir et al., 2006). These productive interactions could facilitate students to draw connections between in-school and out-of-school contexts and see the relevancy of science beyond the science classroom. Additionally, as argued by some researchers (e.g., Warren et al., 2001; Nasir et al., 2006) productive interactions would also promoting student learning.

Productive interactions *among* students should also be facilitated in the science classroom. Students need to be seen as an essential resource in fostering their peers' learning, especially in the area of classroom formative assessments. As discussed in this dissertation, peer feedback needs to be emphasized instead of the peer evaluation techniques (e.g., use of a rating scale to evaluate peer's work, assigning points or grades) currently employed in many classrooms. Students need to be encouraged and supported in providing feedback to their peers and receiving feedback from their peers. This process of engaging students in receiving and providing feedback can help in establishing a community between students and also aid in students' learning.

Implications for Formative Assessment Design and Research

As advocated in the *Frameworks* (NRC, 2012) all aspects of education including assessments should illuminate to students how professionals practice science and engineering. Formative assessment systems should reflect discipline specific activities within the assessment system (e.g., providing feedback, acting on the feedback). The 4R Activity included discipline-specific activities by employing the double-blinded review process to provide feedback to students and provided students with revision opportunities to highlight the iterative nature of knowledge building in science and engineering. Furthermore, this dissertation proposes the use of technological advancements to create innovative formative assessment systems in science that reflect the practices of scientists and engineers. The use of technology can facilitate the creation of learning environments (e.g., double-blinded review) where students are engaged with the practices of scientists and engineers.

With creation and administering of new formative assessment systems, sustained support needs to be offered to teachers to help them in implementing the activity in their classrooms. In the 4R Activity sustained support was provided to teachers through weekly discussions with teachers. In the weekly meetings, aspects of implementation (e.g., supports provided to ELLs, students with IEPs) were discussed with the teachers and their suggestions for improvements were documented. With respect to providing support to students, an informational video on the 4R Activity, its different stages and its applicability to the work of professionals was screened in the classrooms. Furthermore, the author of this dissertation discussed with students the intent of the 4R Activity and answered their questions after the screening of the video. Also, prior to the first 4R Activity an offline practice activity on peer review and revision was administered to help students engage with the peer review and revision process. The practice activity allowed students to engage in a discussion with the teacher on different aspects of peer review and revision.

Beyond the design of new and innovative formative assessments in science future work should also look at achievement gains of students from traditionally underrepresented populations in science and engineering disciplines. While this dissertation examined and reported on the achievement gains of sub-groups as such females, African Americans, English Language Learners and students with IEPs, more work in this area is needed to understand the role of formative assessments in the advancement of learning of certain sub-groups. It is only by doing so will we further the vision of the *Framework* (NRC, 2012) to support *all* students in developing sufficient knowledge of the science and engineering disciplines.

References

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

Biswas, G., Leelawong, K., Schwartz, D., & Vye, N. (2005). Learning by teaching: A new agent paradigm for educational software. *Applied Artificial Intelligence*, *19,* 363-392.

Burstein, J. (2003). The e-rater scoring engine: Automated Essay Scoring with natural language processing. In M. D. Shermis and J. C. Burstein (Eds.), *Automated Essay Scoring: A cross disciplinary approach* (pp. 113–121). Mahwah, NJ: Lawrence Erlbaum Associates.

Black, P., &Wiliam, D. (2004). The Formative Purpose: Assessment Must First Promote Learning. Yearbook of the National Society for the Study of Education, 103: 20–50. doi: 10.1111/j.1744-7984.2004.tb00047.x

Black, P., & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education: Principles, Policy and Practice, 5*(1), 7–73.

Corbett, A. T., Koedinger, K. R., & Anderson, J. R. (1997). Intelligent tutoring systems. In M. Helander, T. K. Landauer, & P. V. In Prabhu (Eds.), *Handbook of human-computer interaction* (pp. 849-874). Amsterdam: North-Holland.

Cameratti-Baeza, C.G. (2016). *Participating in specialized Discourse to learn teaching. A qualitative study of the Ann Arbor Languages partnership* (Unpublished doctoral dissertation). University of Michigan, Ann Arbor. Dissertation in preparation.

De Guerrero, M. C. M., & Villamil, O. S. (2000). Activating the ZPD: Mutual scaffolding in L2 peer revision. *The Modern Language Journal, 84*, 51–68.

Gotwals, A. W., & Songer, N. B. (2009). Reasoning up and down a food chain: Using an assessment framework to investigate students' middle knowledge. *Science Education, 94*(2), 1-23.

Intelligent Essay Tutor (2010). [Automated Scoring Software]. Boulder, Colorado: Pearson's Knowledge Technologies.

Kennedy, C. A., & Wilson, M. (2007). Using Progress Variables to Interpret Student Achievement and Progress (BEAR Technical Report No. 2006-12-01). *Retrieved from the Berkeley Evaluation & Assessment Research Center website: http://bearcenter. berkeley. edu/publications/UsingProgVars_Kennedy_Wilson. pdf.*

Lundstrom, K., & Baker, W. (2009). To give is better than to receive: The benefits of peer review to the reviewer's own writing. *Journal of Second Language Writing*, *18*(1), 30-43.

Martin, J. (2004). Self-regulated learning, social cognitive theory, and agency. Educational Psychologist, 39, 135-145.

Minstrell, J., & van Zee, E. (2003). Using questioning to assess and foster student thinking. *Everyday assessment in the science classroom* (pp. 61-73).Arlington, VA: National Science Teachers Association.

National Research Council. (2012). *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*. Committee on a Conceptual Framework for New K-12 Science Education Standards. Board on Science Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Nasir, N. I. S., Rosebery, A. S., Warren, B., & Lee, C. D. (2006). Learning as a cultural process: Achieving equity through diversity. *The Cambridge handbook of the learning sciences*, 489-504.

National Research Council. (2014). *Developing Assessments for the Next Generation Science Standards.* Committee on Developing Assessments of Science Proficiency in K-12. Board on Testing and Assessment and Board on Science Education, J.W. Pellegrino, M.R. Wilson, J.A. Koenig, and A.S. Beatty, *Editors.* Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Page, E. B. (2003). Project Essay Grade: PEG. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 43–54). Mahwah, NJ: Lawrence Erlbaum Associates.

Rudner, L. M. & Liang, T. (2002). Automated essay scoring using Bayes' theorem. *The Journal of Technology, Learning, and Assessment, 1*(2), 1-22.

Sandoval, W. A., & Millwood, K. A. (2005). The quality of students' use of evidence in written scientific explanations. *Cognition and Instruction, 23*, 23–55.

Teo, A. K. (2006). Social-interactive writing for english language learners. *The CATESOL Journal, 18*, 160–178.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard Univ. Press

Wiliam, D., & Thompson, M. (2007) Integrating assessment with instruction: what will it take to make it work? In C. A. Dwyer (Ed.) *The future of assessment: shaping teaching and learning* (pp. 53-82*).* Mahwah, NJ: Lawrence Erlbaum Associates.

Warren, B., Ballenger, C., Ogonowski, M., Rosebery, A. S., & Hudicourt-Barnes, J. (2001). Rethinking diversity in learning science: The logic of everyday sense-making. *Journal of research in science teaching*, *38*(5), 529-552.

Appendices

Appendix A: Focal Question, NGSK Product Prompt, Assessed NGSK and the Dimensions of Learning Assessed

Focal Question	Prompts for constructing the NGSK Product in the Pretest/Posttest/4R Activity	Assessed NGSK	Science and Engineering Practices, DCI, Crosscutting Concepts from Next Generation Science Standards (NGSS)
1. How do biotic and abiotic factors affect where species lives?	Can grey foxes live in Montana? (Pretest/Posttest) Can red foxes live in Montana? (4R Activity)	Construction of a NGSK product by using a representation (temperature map) and information on a species and its prey's temperature ranges to show that the distribution of organisms is dependent on their environmental interactions both with other living things (biotic) and with nonliving things (abiotic).	Science and Engineering Practices. o Analyzing and Interpreting Data. Analyze and interpret data to provide evidence for phenomena. o Constructing Explanations and Designing Solutions. Construct an explanation that includes qualitative or quantitative relationships between variables that predict(s) and/or describe(s) phenomena. DCI. o LS2.A. Interdependent relationships in ecosystems. Organisms and populations of organisms are dependent on their environmental interactions both with other living things and with nonliving things. Crosscutting Concept. o Cause and Effect. Relationships can be classified as causal or correlational, and correlation does not necessarily imply causation.
2. How has Earth's temperature	If carbon dioxide emissions continue to increase in the future,	Construct a NGSK product by analyzing and interpreting temperature	Science and Engineering Practices o Analyzing and Interpreting Data. Analyze and interpret

changed in the last 150 years?	what do you predict will happen to the Earth's surface temperature? (Pretest/Posttest/4R Activity)	data of the past century to predict the effect of increasing carbon dioxide emissions on Earth's mean surface temperature.	data to provide evidence for phenomena. ○ Constructing Explanations and Designing Solutions. Construct an explanation that includes qualitative or quantitative relationships between variables that predict(s) and/or describe(s) phenomena. DCI. ○ ESS3.D. Global Climate Change. Human activities, such as the release of greenhouse gases from burning fossil fuels are major factors in the current rise in Earth's mean surface temperature. Crosscutting Concept. ○ Cause and Effect. Relationships can be classified as causal or correlational, and correlation does not necessarily imply causation.
3. How will climate change affect species distribution?	Will red squirrels be found in Ohio in the future? (Pretest/Posttest) Will brown squirrels be found in Ohio in the future? (4R Activity)	Construct a NGSK product by using graphical displays of large datasets as evidence to show the cause and effect relationship between organisms and their interaction with nonliving factors.	Science and Engineering Practices. ○ Analyzing and Interpreting Data. Use graphical displays of large data sets to identify temporal and spatial relationships. ○ Constructing Explanations and Designing Solutions. Construct an explanation that includes qualitative or quantitative relationships between variables that predict(s) and/or describe(s) phenomena. DCI. ○ LS2.A. Interdependent Relationships in Ecosystems. Organisms and populations of organisms are dependent on their environmental interactions both with other living things and with nonliving things. Crosscutting Concept. ○ Cause and Effect. Relationships can be classified as causal or correlational, and correlation does not necessarily imply causation.

Appendix B: Holistic Scoring Rubric and its Characteristics

Adapted with permission from "Classroom-based holistic assessment and evidence of middle school students' Next Generation Science Knowledge", by N.B. Songer, S.Z. Zaidi & M.R. Newstadt, in preparation.

Score Level	Description	Characteristics of each level	Distinguishing Characteristics between each level
Level 5 (Proficiency)	**Demonstration of a complete knowledge product that effectively incorporates relevant scientific terminology.** *Elaboration:* Appropriate DCI and CC are used in the explanation/prediction, which are elaborated upon. The explanation/prediction includes a claim supported by relevant evidence along with appropriate reasoning that links the claim and evidence. AND NGSK product effectively incorporates relevant scientific terminology.	• Explicit mention of the source of evidence (e.g., the brown squirrels distribution map) • Response is adequately elaborate. • Additional supporting evidence provided from previous lessons in the curriculum.	
Level 4 (Proficiency)	**Demonstration of a complete knowledge product but lacking effective incorporation of relevant scientific terminology.** *Elaboration:* Appropriate DCI and CC are used in the explanation/prediction, which are elaborated upon. The explanation/prediction includes a claim supported by relevant evidence along with appropriate reasoning that links the claim and evidence. AND NGSK product does not incorporate any relevant scientific terminology, or uses only some of the relevant scientific terminology.	• Explicit mention of the source of evidence (e.g., the brown squirrels distribution map) • Response is adequately elaborate. • Additional supporting evidence provided from previous lessons in the curriculum.	Distinguishing characteristic between level 4 and 5 • In effective use or absence of scientific terminology, such as referring to the trend line as "line on the graph", or, referring to prey as fox's food.
Level 3	**Demonstration of a partially complete knowledge product that is appropriately elaborated upon and effectively incorporates some relevant scientific terminology.** *Elaboration:* Appropriate DCI and/or CC are used in the explanation/prediction, which are elaborated upon. The explanation/prediction is partially complete as one of the three parts of the explanation (claim, reasoning, evidence) is missing or not explicitly stated. AND NGSK product effectively incorporates some relevant scientific terminology.	• If evidence is provided then source of evidence is explicitly stated. • Response is adequately elaborate • Use of some scientific terminology, which is incorporated effectively.	Distinguishing characteristic between level 3 and 4 • Absence or lack of explicitly stating one part of an explanation
Level 2	**Demonstration of a partially complete knowledge product that lacks elaboration and does not incorporate relevant scientific terminology.**	• Presence of a complete claim and accurate evidence/reasoning.	Distinguishing characteristic between level 2 and 3

	Elaboration: Appropriate DCI and/or CC are used in the explanation/prediction, which are not adequately elaborated upon. The explanation/prediction is partially complete as two of the three parts of the explanation (claim, reasoning, evidence) are missing or not explicitly stated. AND NGSK product does not effectively incorporate relevant scientific terminology.	• Response lacks elaboration • Absence of scientific terminology	• Absence of stating the source of evidence or ineffectively stating the source of evidence (e.g., "On a graph, it shows the temperature is rising") • Two parts of the explanation are absent or underspecified. • Absence or minimal use of scientific terminology
Level 1	**Demonstration of an incomplete knowledge product that lacks elaboration and does not incorporate relevant scientific terminology.** *Elaboration:* Only uses appropriate DCI, which is not adequately elaborated upon. AND NGSK product does not effectively incorporate relevant scientific terminology.	• Presence of an accurate claim only. • Claim may be supported by a conjecture instead of evidence. • Response lacks elaboration • Absence of scientific terminology	Distinguishing characteristic between level 1 and 2 • Two parts of the explanation are absent or inaccurate.
Level 0	**Demonstration of an inaccurate knowledge product.** *Elaboration:* Inappropriate DCI or scientifically inaccurate use of the DCI	• Scientifically inaccurate response	Distinguishing characteristic between level 0 and 1 • Inaccurate claim

Appendix C: Peer Review Sheet Used in the 4R Activity

Center for Essential Science

Review Sheet for Explanation/Prediction

	Area to Improve	Suggestions for Improvement
A	Not all parts of the explanation/prediction are present. (One or more of Claim, Reasoning or Evidence is missing)	Please add ___________ (specify missing parts) to your explanation/prediction. In your ________ (specify missing parts) include the following details: ____
B	Explanation/Prediction does not have specific evidence	Please look at information provided (example: maps, graphs) to come up with more specific evidence. In your evidence, include the following details: _____________________________________
C	Explanation/Prediction does not have specific reasoning	Please think about which scientific fact (reasoning) that links your claim and evidence, and add that to your explanation/prediction. In your reasoning, include the following details: ______________________________________
D	Your explanation/prediction needs stronger scientific words	Please think about the scientific words you learned in this lesson and add one or two to your explanation/prediction. Suggested words: ________________
E	Other area of improvement:	Your Suggestion:

Appendix D: Reflection Categories with Supporting Examples

Category	**Example from Student Reflections** **(Direct Quotes)**	**Scientist's Comments (Feedback to the Reviewer & Scientist's Review (wherever applicable)** **(Direct Quotes)**
Reflection on the evaluative statement in the scientist 's feedback to the reviewer		
1. Reflection is only about the evaluative statement made by the scientists on the reviewing skills.	"I learned that I gave good feedback."	**Scientist's feedback to the REVIEWER:** "Good job done in reviewing and good areas selected. In the future, you should provide hints and details of what should be added. Give them example sentence. "
2. Reflection includes that no improvement is required because of the evaluative statement made by the scientist.	"The scienstist said that I did good so I don't need to improv."	**Scientist's feedback to the REVIEWER:** "Good attempt made in reviewing. In the future, you should provide more details of what the author should be looking at on the map. Give them examples and hints."
Reflection on the scientist's feedback to the reviewer		
3. **Verbatim:** Includes aspects of what the scientists said using direct quotes from the scientist's feedback.	"What i learned is to provide examples of what sentences need to be corrected."	**Scientist's feedback to the REVIEWER:** "Good attempt made in reviewing. In the future, you can provide examples of what sentences need to be corrected. I am not sure what you wanted the author to revise and where those revisions need to be made."
4. **Paraphrased** from scientist's feedback	"One thing that I learned from reading the scientist's review is that I could tell them the reason for making the changes that I suggested. This will help	**Scientist's feedback to the REVIEWER:** "Good job done in reviewing. You provided good suggestions and example sentences. In the future, you can describe why the changes you suggested

	<u>them understand why my changes will make their prediction better.</u>"	<u>will make their prediction stronger,</u> such as reminding them what evidence is or what reasoning is."
5. **Elaborated reflection** on the scientist's feedback to the reviewer	"<u>I need to describe more when I'm reviewing because I kind of just say you need to fix "that" but I think I need to go into more detailed and say you need to fix what your saying about 1900-2010 because my peer might not understand what I'm saying in the review.</u>"	**Scientist's feedback to the REVIEWER:** "Good job done in reviewing. <u>In the future, try to give more details in your suggestions when you ask the author to make specific changes. Also make sure that you indicate where the change needs to be made. Give them example sentences of what they can add or hints about where they can look for the information to be added</u>. See below to get an idea…"
Reflection on scientist's feedback to the reviewer and scientist's review		
6. Reflection includes (*a) feedback given by the scientist to the reviewer* and (b) <u>referencing the scientist's review.</u>	"I learned <u>next time instead of saying something small, like saying that they need to separate the different parts of the "My Prediction," I should focuss on the more important things,</u> <u>such as the example the scientist gave about improving the reasoning:</u> "<u>In your reasoning, you need to add more details such as "the temperature in Ohio will be higher than the range of temperatures that the brown squirrel lives in</u>". *Another thing I learned is that in the future I should think hard and help my peer with giving them more feedback such as scientific words that they could add.* I will keep this in mind."	**Scientist's feedback to the REVIEWER:** "Good job done in reviewing. *In the future, try to look for other areas of improvement, such as words that can be added.* See below to get more ideas for future reviewing. **Scientist's Review for the AUTHOR:** "Suggested Steps for Revision: 1. <u>I don't think that you necessarily need to separate each section, like your peer suggested, as you have begun each section with "my reasoning/evidence...". Doing it this way is also acceptable.</u> 2. <u>In your reasoning, you need to add more details such as "the temperature in Ohio will be higher than the range of temperatures that the brown</u>

		squirrel lives in" instead of saying "too hot for the brown squirrel to live in Ohio"…."
7. Reflection includes (**a) self-assessment of their peer reviewing skills,** *(b) feedback given by scientists* (c) referencing the scientist's review	"**Some things I learned from this exercise is that I shouldn't focus on the small pieces and get the big picture, such as not being focused on grammar so much.** I should begin with selecting more important areas of improvement. *In the future, I could suggest more scientific vocabulary to use to make my peer's prediction stronger.* I really liked that in the scientist's review, the scientist gave good hints and examples and included little bits and parts from my peer's prediction to show the areas they need to improve. This can help my peer because they can know exactly where to put the revisions that are being suggested."	**Scientist's feedback to the REVIEWER:** " Good job done in reviewing. *In the future, try to suggest other areas of improvement also, such as scientific words that can be added."* **Scientist's Review for the AUTHOR:** "Suggested Steps for Revision: … 2. In the reasoning, describe WHY the brown squirrel will not be found in Ohio. You wrote, "they cannot live there because it will be hot". You need to provide details about increase in temperatures in the future (and state where you got the information from) and WHY this will affect the squirrel. (Hint: remember that animals can only live in places as long as the abiotic conditions (example: temperature) are suitable for them in the place). 3. In your evidence, you refer to the map as "map from start". Instead, refer to the map as "distribution map of brown squirrels" as the map provided shows where brown squirrels are found, therefore this map is a distribution map of brown squirrels."

Appendix E: Examples of "Additional Aspects" in Reflection

Additional Aspects	**Examples from Student Reflections (Direct Quotes)**
a. Reflecting on how they would have improved upon this round of review.	"I learned that it is important to give specific peer reviews. In the scientist's review she gave specific examples to my peer but I don't really feel like I gave my peer many points and specific areas to fix. I think that if I did that it would have been more helpful for my peer. I also think some of the advice I gave to my peer was very obvious and I could have given here better advice specific areas in my peer's text such as when my peer said, "As the Carbon increases and rises into the atmosphere the global surface temperature will increase." <u>I could have asked my peer to explain why carbon dioxide causes the global surface temperature to rise."</u>
b. Including a reflection on authoring the NGSK product	<u>"The first thing I learned is that it is important that I read the information provided more carefully and think about how to use it when I write my own explanation. I did not do that and therefore I had many missing things.</u> I think I need to add more examples on how to implement my recommendations so that my peer is clear on why and how to add the suggestions I give."
c. Linking authoring and reviewing NGSK products	"I have learned that I need to go into detail and make it clear what they need to add. It is important to give examples so they understand more clearly what I am asking them and let them know how to include those examples. When I ask to add words I should tell them what words to add and why. <u>I forget to include these things in my own explanation so I should also remember these things when I write my own explanation</u>. <u>I think that once I do that I can do better reviews also."</u>
d. Reflecting on the purpose of the 4R activity	<u>"From my understanding the 4R Activity was avout learing to be a better scientist and learning how to review better and write better explanations and predictions. This will be very helpful later on in life and I think I learned how to do so. I think the reviews taught me to not only fix my mistakes but also to write better. It was helpful to not only write a explanation but to also review others and get feedback</u>…"

Appendix F: Semi-Structured Interview Protocol

Introduction with the Student and Setting up
(1) Inform them of what the interview is about.

Script: I am planning on using the 4R activity again with other students, hopefully from all over the country. I would really appreciate hearing your thoughts about the 4R activity. For example, I would like to know what suggestions you have for improving the activity. What would you suggest I do to make the activity a better learning experience for other students?

Recap the 4R Activity.

Relevant to their life inside and outside school
1.Have you used something similar to the 4R Activity in science class before?
Clarification: Have you done any of the 4Rs?
  a. When?
  b. Are you able to describe what you did/that situation (or describe the example you're thinking of)? How did that go?
2. Have you used something similar to the 4Rs in other classes?
  a. What class?
  b. Are you able to describe what you did (or describe the example you're thinking of)? What did you do in it, how did that go?

If Response to (2) == No, then.
a. Do you think it will be beneficial/useful to use the 4Rs in a class other than science?
  a. Why or why not?
  b. Can you describe an example?
  c. What class would it be?
  d. How would you use it?
  e. What about other parts on the 4R Activity?

Ask them something that they do for leisure: sports, music, dance, video games, fine art, reality shows.
X=leisure
Y= <Professional related to the activity listed (e.g. dancer, musician, football player, video game developers)>
3. Can you think of using the 4R Activity when you do X?
  a. Why or why not?

Use of the 4R Activity by Professionals
1. Do you know of ways or can you think of ways in which a Y professional uses some or all of the 4R's?
  Clarification: Similar to some/any of the R's?
2. Elaboration on their answer to (1).

If “ No” for (1) then,

a. Why do you think that?

3. How about other professionals?
4. How about scientists or engineers? Do you think they use any part of the 4Rs in their work?
5. Ask for elaboration (possible questions listed below)

a. Ask them how do they know that scientists and/or engineers do this?

Prompt: have you seen it on TV, or Internet, you know of a person, you read it somewhere, your friend told you…

b. Ask them if they can give an example of when scientists or engineers use 4R.

c. Ask them why they think scientists or engineers use the 4Rs (i.e., why they think the 4Rs might be important to scientists’ or engineers’ work).

d. If “No” to (4), ask them why they think the 4R process is different from (or “doesn’t apply to”) what scientists and engineers do

---

Self report on student perceptions of learning

1. When I talk about the 4R Activity, what comes to your mind about what you learned from participating in it? It could be anything. If you feel you didn’t learn anything in particular you can also say that, because that is very helpful to me too.

[If the student struggles, then say, “For example” and list one of your prompts.]
Prompt: learned <content>, learned how to make a stronger explanation, peer reviewing, using review to revise, thinking about how to give a better review in the?

2. Ask for elaboration, if not done already.

3. Ask them that if there were anything else they would want to talk about with respect to what they learned?

*4. What did you think about being able to revise your explanation/prediction?

a. Was your revised explanation different than your original explanation?

b. Can you talk a little bit more about it?

*5. How was it to give review to your peer?

a. Was giving peer review useful to you in anyway?

b. Can you give an example?

*6. When you were doing the reflection, what did you think about what the scientist had to say about your reviewing skills?

a. Was it useful to you in any way later on?

b. Can you give an example?

---

*Note. All students do not answer 4, 5 and 6. These questions also served as elaboration questions for some participants that mentioned one or some of the above three things as what they learned from the 4R Activity.

Student Suggestions for Improvement

(1) What did you think of the 4R activity?
Prompt: Did you like it, didn't care much about it, did not like it.

(2) Can you talk about what you liked/didn't like about it?

(3) What would have made it better for you?
(Build this question on what they said they didn't like/like about it)
Prompt: give examples such as, (1) less writing, (2) it was too long/too short, (3) the type of feedback you received…
(4) Were the directions to the activity clear?
(5) Do you have any ideas, or suggestions to improve this activity?

Appendix G: Examples of Students Self-Report on Learning Tied to Future Aspirations

Career/Future Aspiration	**Student Responses to Learning from Participating in the 4R Activity**
Teacher/Professor	Student 1: "That [giving peer review] would be very helpful to me later on because I want to become a teacher or a professor I haven't decided on that right now, but definitely one of those things. And learning how to give feedback is what they have to do all the time, so I feel that I learned it now already and now I know that you have to really read carefully and think of how to help the other person, instead of just trying to not take it [giving feedback] seriously and finish it quickly."
Scientist	Student 2: "I liked it because it will help me become a better scientist in the future and I learned more things that scientists do, like write clearly for other people to read and receive review and revise and also give reviews. Oh also reflect on how to be better on everything like writing and giving better reviews."
Engineer College Preparedness	Student 6: "I liked it [using feedback to revise] and I think that it will really help me in future scenarios when I become an engineer and have to design over and over and get it refined or even other scenarios like this, in other classes, in college, so if I am ever faced with something that I have to review or respond to something or reflect on something, I think that that'll be stronger now that I've done this."
College Preparedness	Student 19: "I learned how to use the feedback. That's very important in college you know. My brother is in college, and he say it is hard because he writes papers and its like feedback on papers and no exams. And I told him that I got feedback and he was surprised coz he didn't get any in science when he was in school."
Class Participation	Student 4: "Using the words [scientific terminology] will be helpful because-- then I can talk in class more about science and say what I mean clearly to others."