2015

Do teacher judgment accuracy and teacher feedback predict student achievement in elementary and middle-school science?

Mason Albert Kuhn
University of Northern Iowa

DO TEACHER JUDGMENT ACCURACY AND TEACHER FEEDBACK PREDICT STUDENT ACHIEVEMENT IN ELEMENTARY AND MIDDLE-SCHOOL SCIENCE?

An Abstract of a Dissertation

Submitted

in Partial Fulfillment

of the Requirements for the Degree

Doctor of Education

Approved:

_____

Dr. Elana Joram, Committee Chair

_____

Dr. Kavita Dhanwada
Interim Dean of the Graduate College

Mason Albert Kuhn

University of Northern Iowa

December 2015

ABSTRACT

The quest to find the most effective approach to teach science has challenged teachers and scholars for decades (Ford, 2012; Kuhn, 2010a; Duschl, Schweingruber, & Shouse, 2007). Unfortunately, the recent push for quality science education in schools has not yet resulted in higher achievement on standardized science assessments in the United States. The increased focus on science in school and the lack of growth on national and international standardized assessment has led to reformed-based policies, such as the Next Generation Science Standards (NGSS), that ask teachers to teach in a way that emulates the practices of scientific inquiry, including argument-based inquiry.

Current literature suggests that using dialogic feedback is an effective way to promote scientific argumentation in classrooms (Chin, 2007), yet we know very little about what teachers need to know in order to provide effective dialogic feedback in an argument-based inquiry classroom. A type of knowledge that may be essential for teachers to access and use, especially when providing feedback, is one about which little is known: teachers' task-specific knowledge of learners' understanding of science concepts. In this study, I used a method for capturing this type of teacher knowledge and explored its relationship to instructional decision-making and, ultimately, to student achievement in science.

Specifically, I examined the links among teachers' knowledge of students' understanding of scientific concepts, the type of feedback teachers gave, and students' science achievement outcomes in classrooms that encourage argument-based inquiry. Teachers' knowledge of science learners was measured using a teacher judgment

accuracy task. Teachers in the study predicted how well their students would perform on specific items on a science assessment. To tap into teachers' instructional decision making, the dialogic feedback they provided in a video-taped science lesson was measured using an observational coding scheme developed for this study.

Thirty-three third-through eighth-grade teachers in two moderate-size school districts in the mid-west United States participated in the study. Hierarchical multiple regression analyses were conducted to examine the relationships among the variables, and dialogic feedback was found to be a significant predictor of student outcomes on the science section of the Iowa Assessments, accounting for a large amount of the variance in those scores. Judgment accuracy was not a significant predictor of student outcomes and accounted for much less variance in the scores than did dialogic feedback. Interaction effects were investigated through separate moderation and mediation analyses, and neither produced statistically significant results.

Results of this study are discussed in terms of their potential to provide insights into teacher knowledge of learners and an instructional decision-making practice (dialogic feedback). These results have the potential to add to our understanding of teachers' knowledge of learners, its relationship to instructional decision making, and its role in student achievement. Implications for both preservice teacher education and inservice professional development are discussed.

DO TEACHER JUDGMENT ACCURACY AND TEACHER FEEDBACK PREDICT STUDENT ACHIEVEMENT IN ELEMENTARY AND MIDDLE-SCHOOL SCIENCE?

A Dissertation

Submitted

in Partial Fulfillment

of the Requirements for the Degree

Doctor of Education

Approved:

_____

Dr. Elana Joram, Chair

_____

Dr. Jill Uhlenberg, Committee Member

_____

Dr. Anthony Gabriele, Committee Member

_____

Dr. Benjamin Forsyth, Committee Member

_____

Dr. Dawn Del Carlo, Committee Member

Mason Albert Kuhn

University of Northern Iowa

December 2015

ACKNOWLEDGEMENTS

I would like to thank the esteemed members of my committee for their expertise and dedication to this project. It is amazing to think how far I have come in such a short period of time, none of which would have been possible without your guidance. You have given me the tools to succeed as an academic; and for that I am forever grateful.

I would like to thank Dr. Benjamin Forsyth for his exemplary work on the committee and for his thought provoking courses *Inquiry* and *Context of Contemporary Education.* Those courses unearthed many of the ideas that set the stage for this study. I am not sure if I would have had the theoretical base I needed for Chapter 2 without the content and the way that you challenged me as a student.

In addition I would like to thank Dr. Dawn Del Carlo for helping me develop the dialogic teacher observation tool and the questions for the prediction task. I think the feedback tool will be the focus of my research agenda over the next few years and I greatly appreciate your dedication to seeing it come to fruition. I look forward to collaborating with you as I fine tune the feedback tool and attempt to publish.

I also would like to thank Dr. Jill Uhlenberg for your guidance, encouragement, and attention to detail. My final document was spotless because of your demand for an excellent product. I look forward to working with you in the future.

Next, I would like to thank Dr. Anthony Gabriele for pushing me throughout the entire process. It was only about a year ago when we sat down and mapped out a potential diagram for this study. When I left the meeting I was so confused I had a

headache, but your guidance, via journal articles, books, and forcing me to find answers helped me slowly find my way. In the end the experience will make me a better scholar.

Finally, I would like to thank my chair, Dr. Elana Joram. Your combination of encouragement and demand for high-quality work is the primary reason I was able to complete a research project that I am very proud of. I am forever grateful to your dedication to the countless edits made to this document; each one pushing my limits just a little farther. I think I could actually feel dendrites dancing through my head after some of our thought-provoking meetings throughout this endeavor. Thank You.

In addition to the committee members, I would like to thank a few other people who helped me complete this project. First, I would like to thank Dr. Mark McDermott for his input on various aspects of the study, but also for his tutelage as I transitioned from a classroom teacher, to a leader in the district, and now as a scholar. I doubt I would have filled out the application to the doctorate program if it wasn't for those years when I was allowed to help develop professional development for teachers through the grant work at Wartburg.

I would also like to thank Thomas Fields for you diligent and timely work helping me code the teacher videos for the dialogic feedback measure. I know you had better things to do on those late nights and weekends so I appreciate your commitment to this study.

I also would like to thank Mark, Denise, and Alex Kuhn; my amazing parents and brother who have been incredibly supportive during this entire process. Their care and belief in me are why I have made it where I am today.

Finally, I want to thank the most important people in my life; my wife Lori, two sons, Tristan and Jake, and my daughter Julia. These last few years have been extremely trying due to the amount of time this project has taken. I have so much joy in my heart knowing that I won't have to tell you "Sorry, I can't play right now; daddy has to do work" as often as I have had to the last few years.

There is no possible way I could have completed the program without the love and support from my best friend. All the night classes, early-morning writing, and weekends at the library would have been impossible without your sacrifice. This last year has been really tough, especially since you have been pregnant, but you have put our family in front of everything. I love you.

TABLE OF CONTENTS

PAGE

LIST OF TABLES

TABLE PAGE

LIST OF FIGURES

FIGURE PAGE

CHAPTER 1

INTRODUCTION

The quest to find the most effective approach to teach science has challenged teachers and scholars for decades (Duschl, Schweingruber, & Shouse, 2007; Ford, 2010; Kuhn, 2010a;). With the United States moving from a manual-labor to a technology-based work force, effective science education is a high priority for both schools and the private sector. According to the Department of Labor's Bureau of Labor Statistics, Science, Technology, Engineering, and Math (STEM) fields have the greatest potential for job growth in the 21st century, and recipients of those jobs will earn higher wages than other fields of work (Lacey & Wright, 2009). During his 2011 State of the Union Address, President Barack Obama called for funding for "100,000 new STEM teachers over the next ten years" (State of the Union, 2011). The President's focus on STEM teaching has prompted dozens of non-profit organizations to band together to form "100Kin10," an organization that hopes to improve the number of qualified science teachers in the United States and build a movement to improve STEM education.

Preparing students for careers in science is a potential outcome of science teaching and is one rationale behind the intense emphasis on quality science teaching in school curricula. High school science is traditionally seen as a route to higher education and career choices for those students interested in the field. From this viewpoint, the science may serve two purposes; first, to prepare those who will enter a scientific career with the knowledge and skills needed; and second, to help students to score sufficiently

on college entrance examinations and thus improve their chances of success in those courses.

A second rationale for the current emphasis on quality science teaching is to develop a scientifically literate population. From this viewpoint, the goal of school science is to provide an education that will be of value to students over a lifetime, regardless of their careers, and thus highlights the scientific needs of all students. Scientifically literate individuals tend to be described as being able to solve problems for health and survival, understand complex issues in a civic way, participate more fully in debate and decision making, as well as being motivated to know about science as a human endeavor (Bybee & McInerney, 1995; Deboer, 2000). With many of the world's 21st century political issues becoming more science-based (e.g. climate change, pollution, medical decision making) the importance of having a scientifically-literate population has never been greater.

Unfortunately, the push for quality science education in schools over the last few decades has not produced higher achievement on national science assessment measures. According to the National Assessment of Educational Progress (NAEP) science scores revealed slight, but not statistically significant, gains from 2009 – 2011 (Martin, Mullis, & Foy, 2012). The Trends in International Math and Science Study (TIMSS) report claims that there is no measurable difference among the average science scores of U.S. students in grade 4 in 1995 (542), 2007 (539), and 2011 (544). Similarly, there is no measurable difference between the average science scores of U.S. students in grade 8 in 2007 (520) and 2011 (525; Martin et al., 2012). Data gathered from the top performing

students in the United States is not encouraging either: only 10% of U.S. 8th graders reach the advanced science level on the TIMSS exam. In contrast, 25% of 8th graders in China and 32% of 8th graders in Singapore reach the advanced science level on the TIMSS (Martin et al., 2012).

Despite the stagnant science assessment results of the NAEP and TIMSS as an aggregate, some science education researchers have reported an improvement of student achievement on standardized science assessments after teachers have developed an enhanced understanding of a targeted teaching approach and/or improved instructional decision making practices (Hand, 2009; Kahle, Meece, & Scantlebury, 2000; Marx et al., 2004; Scruggs, Mastropieri, Bakken, & Brigham, 1993). Because of the pressing need to enhance U.S. students' understanding of science and their performance on standardized assessments a detailed look into aspects of quality science teaching is needed.

This study examined two variables that may be related to the effectiveness of science teaching. It focused on a form of teacher knowledge, teacher judgment accuracy, as well as instructional decision making, and examined their predictive relationships with student outcomes on a standardized assessment. An overview of the problem will be presented next, followed by a detailed discussion of the two predictor variables, judgment accuracy and instructional decision making.

Theoretical Framework

The act of science is a practice grounded in argumentation (Anderson, 2002). Scientific argumentation has been described as an attempt to validate or refute a claim on the basis of reasons that reflect the values of the scientific community (Norris, Philips, & Osborne, 2007). During the processes of scientific inquiry, scientists make claims based on observable evidence, and clarify, with explanation of the evidence as related to the claims (Kuhn, 2010a). Other scientists often make rebuttals, pointing to other evidence that counters the evidence from the previous claim. The key aspects of this interaction are that any scientific claims coming from an investigation must be based on empirical evidence, and that evidence must be justified as connected to the claims. The dialogical nature of scientific discourse practiced among scientists is a key reason reformed-based science standards established by the National Research Council (NRC) and the Next Generation Science Standards (NGSS) suggest teachers employ classroom practices that promote students constructing arguments from evidence and critiquing each other's claims through negotiation (NGSS Lead States, 2013). The practice of scientific argumentation, practiced by actual scientists, has been suggested as an activity students in science classrooms should emulate (Ford, 2012; Hand, 2007; Kuhn & Crowell, 2011; National Research Council, 2012). The role of argument as a framework for classroom instruction will be discussed in the next section.

Argument-Based Inquiry

Science education plays a vital role in preparing students for various aspects of their future lives: thinking logically and critically, making decisions involving scientific information both personally and as active citizens and, for some, pursuing a career in science (*Benchmarks for Science Literacy*, 1993; Duschl et al., 2007; NRC, 2012). Teachers who educate students with these goals in mind, place a special emphasis on teaching the skills of inquiry to students. Learning through inquiry involves the skills needed to ask questions, conduct investigations, generate data, create models, interpret evidence from first-hand investigations, and make evidence-based claims (NRC, 2012). If taught well, the process of inquiry asks students to engage in critical thinking, interpret data, and to consider alternative explanations of evidence (Ford, 2012; Sandoval & Reiser, 2004).

A specific type of inquiry that asks teachers to explore the dialogic interactions of the process of inquiry is argument-based inquiry. Asking students to construct arguments from evidence has been an extensively supported goal in national-level science education policy (Duschl et al., 2007; NGSS Lead States, 2013; NRC, 2012). These policies have asked teachers to promote classroom practices that move beyond experiments and investigations, and towards practicing science argumentation. According to the NGSS "Students should engage in the practices of asking questions, planning and carrying out investigations, analyzing and interpreting data, constructing explanations, and engaging in argument from evidence" (NGSS Lead States, p. 49). Research on how to enact reform-based teaching practices most effectively has suggested that teachers embrace a

role that promotes autonomous learning practices, like negotiating evidence with peers, and moves away from authoritative, or directive lecture (Hargreaves, 2013).

Asking students to engage in explanations of scientific phenomena through argumentation creates opportunities for students to engage in multiple aspects of scientific inquiry while building their science knowledge. When students participate in scientific argumentation they are provided with a context and a foundation for the process skills of inquiry. In addition, due to the nature of argumentation, students necessarily practice the critical thinking skills that are vital to inquiry, as they need to evaluate evidence and critique alternative explanations (Kuhn & Crowell, 2011; Hand, 2009; Ford, 2012). As students engage in the process of critiquing each other's claims, the act of communicating and justifying explanations plays a central role in their inquiry, underscoring key aspects of the nature of science.

Teachers who use instructional decision making practices aligned with the theoretical framework of argument-based inquiry have the potential to practice the skills required to think critically and to address the two previously mentioned rationales for the emphasis on quality science teaching being taught in school. First, skills of scientific argumentation align with the commonly held notion that scientific claims are empirical, tentative, and negotiated by the most rigorous standards of peer-review until accepted by the scientific community (Bricker & Bell, 2008). Shaping a student's epistemological and ontological views of scientific knowledge construction may help equip them with the skills required in a profession in the field of science. Second, an understanding of how

the scientific community determines what is accepted as scientifically viable would increase the scientific literacy of the population as a whole.

Engaging students in the act of scientific discourse requires teachers to structure their pedagogy from a learning theory that allows students to build their own understanding and knowledge of the world, through experience and reflecting on those experiences. The learning theories of constructivism and social-constructivism fit the model described above and will be discussed in further detail in the following sections.

Constructivism

Constructivism is a theory of epistemology that argues that human learning occurs through an active process in which people construct their own subjective representations of reality through their experiences (Reigeluth, 1999). There are two major strands of the constructivist perspective: cognitive constructivism and social constructivism. Cognitive constructivism suggests that ideas are constructed by individuals through a personal process, in contrast to social constructivism whereby ideas are constructed through interactions with other individuals such as teachers and other peers (Powell & Kalina, 2009). Whereas the two perspectives differ in their emphases, they share many common perspectives about teaching and learning (Jonassen, 1999). proposed several characteristics that both cognitive and social constructivist learning environments share; specifically, they, "… allow for multiple representations of content, emphasize knowledge construction instead of knowledge reproduction, emphasize authentic tasks, provide learning environments such as real-world settings or case-based learning,

encourage thoughtful reflection on experience, and allow collaborative knowledge construction through social negotiation" (p. 35). The framework of this study is based on social constructivism which will be discussed further in the next section.

Social-Constructivism

Social constructivism, developed by Vygotsky, emphasizes the collaborative nature of learning. Vygotsky (1978) rejected the assumption of cognitivists such as Piaget who separated learning from its social context. Vygotsky, instead, argued that all cognitive functions originate in, and must therefore be explained as products of social interactions. According to his view, learning is not simply the assimilation and accommodation of new knowledge by the learner; but is rather the process by which learners are integrated into a knowledge community.

Vygotsky believed that culture gives the child the cognitive tools required for development (Scott, 2008). Adults, such as parents and teachers, are channels for the tools of the culture, which include cultural history, social context, and language (Reigeluth, 1999). The major theme of Vygotsky's theoretical framework is that social interaction plays a fundamental role in the development of cognition. Vygotsky (1978) stated that: "Every function in the child's cultural development appears twice: first, on the social level, and later, on the individual level; first, between people (interpsychological) and then inside the child (intrapsychological)" (p. 57). The interpsychological plane Vygotsky discussed suggests that language can be used as a tool for learning (Scott, 2008).

Vygotsky's view of knowledge construction parallels the practice of the scientific community and their acceptance of what is considered a viable scientific claim. Similar to a child negotiating a new concept with peers on the interpsychological plane, actual scientists discuss hypotheses within their respective research circles and argue about what ideas warrant further investigation. Once negotiated on a social plane, scientists internalize the interaction, which leads to further discourse based on the interpretation by the scientist. Teachers who use pedagogy aligned to argument-based inquiry encourage students to mimic similar actions in the classroom.

Thus far, the introduction to the study has reviewed the nature of science as an act of constructing and critiquing scientific claims through a form of argumentation, described a pedagogical framework that uses student dialogue as a tool for learning, provided a description of a learning theory within which the pedagogical approach is situated, and explained how this framework addresses the two rationales for why science is taught in school. In the following sections, the two predictor variables that were used in the study, judgment accuracy and instructional decision making, will emerge from reviewed literature as a type of teacher knowledge and a factor related to quality science teaching, respectively.

Teacher Knowledge

The notion that teachers possess different types of knowledge, and that having mastery of these diverse knowledge bases is required for effective teaching, has been studied extensively over that last few decades. A typology of these knowledge bases was

put forth by Shulman (1986, p. 8) when he described a framework for Pedagogical Content Knowledge (PCK):

1. Content Knowledge
2. General Pedagogical Knowledge
3. Curriculum Knowledge
4. Knowledge of Learners
5. Knowledge of Educational Contexts
6. Knowledge of Educational Ends
7. Pedagogical Content Knowledge (PCK)

Shulman introduced PCK as teachers' "own special form of professional understanding" (Shulman, 1987, p. 8). Shulman (1987) claimed that the emphases on teachers' subject knowledge and pedagogy were being treated as commonly exclusive areas in education research. The practical consequence of this exclusion was the teacher education programs in which a focus on either subject matter or pedagogy dominated. To address this dichotomy, he introduced PCK as a way of bridging content knowledge and pedagogical knowledge. Shulman (1986) acknowledged that much is known about how teachers manage their classrooms, organize activities, allocate time, structure assignments, ascribe praise and blame, formulate the levels of their questions, plan lessons, and judge general student understanding. What was missing from the research were "questions about the content of the lessons taught, questions asked, and the

explanations offered" (Shulman, 1986, p. 7). Shulman's identification of pedagogical content knowledge spawned a shift in emphasis among researchers to studying PCK and its relationship to effective teaching (e.g. Hill, Ball, & Schilling, 2008).

However, the fourth type of teacher knowledge Shulman (1986) identified, knowledge of learners, has not received as much attention as PCK. Little research has been conducted on teachers' knowledge of their own students, yet it may also be critical for effective teaching. Previous research on knowledge of learners has come in the form of creating PCK models that include some or all of Shulman's knowledge bases. For example, Park and Chen (2012) explored the nature of the integration of five components of PCK (Orientations toward Teaching Science, Knowledge of Student Understanding, Knowledge of Instructional Strategies and Representations, Knowledge of Science Curriculum, and Knowledge of Assessment of Science Learning) by tracking the development of each in a small group of expert science teachers. Park and Chen (2012) found that the most common pattern across the teachers' PCK Maps was Knowledge of Student Understanding and Knowledge of Instructional Strategies and Representations. The researchers suggested that the teachers' understanding of student understanding and corresponding teaching strategies were the two variables that were the most influential in moderating classroom instruction (Park & Chen, 2012). Previous research on PCK variables also suggests that teachers' knowledge of student understanding is critical to the development of other PCK variables. Clermont, Krajcik, and Borko (1993) and Loughran, Berry, and Mulhall (2012) indicated that teachers' familiarity of student understanding such as presumptions, learning complications, and reasoning types in a

particular domain enabled the development of their PCK. Although Park and Chen (2012) subsume knowledge of learners under PCK, in the proposed study, I return to Shulman's original typology and examine knowledge of learners as a distinct form of teacher knowledge.

The Role of Knowledge of Learner in Argument-Based Inquiry

Reform-based science-education theoretical frameworks (e.g., argument-based inquiry, argument-driven inquiry, discovery learning, student-driven inquiry) and teaching approaches (e.g., the *Science Writing Heuristic*, the *Five Es*) suggest that K-12 students should emulate the practices of actual scientists. Some science education scholars have suggested that replicating the vocation of science in a classroom may not be feasible (Ford, 2012). Nonetheless, nearly all agree that teaching the skills of questioning, inquiry, data evaluation, generating claims, and making conclusions in the context of science are all worthwhile endeavors (Ford, 2010; Ford, 2012; Green & Luke, 2006, Hand, 2007; Hand, 2009; Hanuscin, 2013; Kuhn & Crowell, 2011; NRC, 2012). These interactions, that help students construct claims based on evidence, is the primary aspect of interest when attempting to replicate how actual scientists make rational decisions.

Implementing the skills of inquiry in students with fidelity requires teachers to acquire pedagogical tools that stretch beyond being able to deliver a traditional didactic lecture. Over the last half-century, most research in science education has focused on facilitative teacher-dialogue (Duschl et al., 2007; Kuhn, 2010b; Sadler, 2006), student-

centered inquiry (Bybee et al. 1989; Hand, 2007) conceptual change (Hewson, 1992; Hewson & Hewson, 1984; Posner, Strike, Hewson, & Gertzog, 1982) and how students develop scientific reasoning skills (Brown & Wilson, 2006). All of these research efforts embrace a constructivist learning approach based on the belief that learning occurs as learners are involved in a development of meaning making and knowledge construction in contrast to passively receiving information.

A teacher applying an argument-based inquiry approach in his or her classroom would, in theory, allow students to search for answers to questions and produce an explanation based on their understanding of the problem at the time. We can conjecture that if the teachers acquire knowledge of their learners, they would then provide feedback and moderate future lessons based on their knowledge of their students' understanding. Teachers who are experts in using this type of teaching in a science classroom should have extensive knowledge of how their students reason through problems, the students' ability to construct a scientifically valid claim, and their overall understanding of the concepts taught (Bruner, 1961; Tobin, 1993). Most science education researchers would agree that a strong knowledge base of student understanding is critical for quality constructivist-based science teaching, but no research could be located that has empirically investigated this type of teacher knowledge. Further, teachers simply acquiring knowledge of student understanding is likely not a determining factor for student success in a classroom; how teachers utilize this information and adjust their instructional decision making may be additional critical factors for promoting academic success.

Teacher Acquisition of Knowledge of Learners

A synthesis of two meta-analyses of inquiry teaching (Hattie, 2009) found that an inquiry-based approach resulted in improved student performance in science classrooms. Notably, inquiry teaching increased the amount of time students spent in labs, negotiating, and knowledge construction, and decreased the amount of teacher-led discussions in classrooms. Hattie (2009) claimed that these instructional practices improved student retention on assessments and improved students' critical thinking skills.

The role of the teacher in argument-based inquiry classrooms, which aligns pedagogically with inquiry-based classrooms, has been described similarly by Sawyer (2006), Hand (2007), Perkins, (2009), Ford (2010), Duschl et al. (2007), and Jonassen, (1999) as: modeling and nurturing the development of habits of minds, encouraging divergent thinking that leads to more questions from the students, valuing and encouraging student responses, effectively exploring the causes and appropriately guiding the learner when their responses convey misconceptions, making student assessment an ongoing part of the facilitation of the learning process, and requiring students to express their understanding of the content and asking them to provide evidence to validate their claims.

The literature on the argument-based inquiry philosophy suggests that effective teachers possess a high level of information about the knowledge of learners because student ideas drive instruction and constantly need to be challenged and revised as they proceed through the learning process. Thus, argument-based inquiry presumes that

teachers already have high levels of knowledge of learners. However, these assumptions have not yet been validated through quantitative research. This study was the first of its kind to attempt to quantify knowledge of learners in science teachers and to evaluate its predictive relationship to student outcomes. Knowledge of learners in the study is viewed as a tool used to guide classroom instruction and lesson planning. To capture a quantifiable measure of knowledge of learners, a judgment accuracy task, described below, was used.

Teacher Judgment Accuracy

Teacher judgment accuracy, which is discussed at length in Chapter 2, has been described as an ability to judge student characteristics correctly and is part of a broader set of skills known as "diagnostic competence" (Artlet & Raush, 2014). Judgment accuracy has also been referred to as a teacher's ability to accurately predict task demands of their students (Anders, Brunner, & Krauss, 2011). Judgment accuracy tasks are considered a way to measure teachers' knowledge of learners because these tasks require teachers to predict student outcomes on assessments. To accurately predict how their students will perform on assessments, teachers would need to acquire knowledge of learners.

The type of judgment accuracy task utilized in this study was the item-specific hit-rate task. The item-specific hit-rate requires a teacher to predict the actual items a student would answer correctly on an assessment. Hit-rate tasks are interesting because the level of understanding a teacher must have about their students' understanding of the

concepts taught must be high. Karing, Pfost, and Artelt (2013) noted that hit rate tasks use "an indicator of all the available information of the judgment of individual students' performance on individual tasks (p. 280)." Thus, the task-specific hit rate takes into account how accurate teachers are at predicting student outcomes on exact items on an assessment, giving a more detailed report of the teachers' understanding of specific skills the student has mastered (Artelt & Rausch, 2014).

The item-specific hit rate was chosen from a range of tasks used to measure judgment accuracy because it examines teachers' knowledge of learners at the most specific level. Access to this information can provide insights into how well teachers can predict their students' comprehension of the content taught in the classroom, and the degree to which they can predict each individual student's level of understanding. A framework for how the task specific hit rate relates to the constructs previously discussed is presented below:

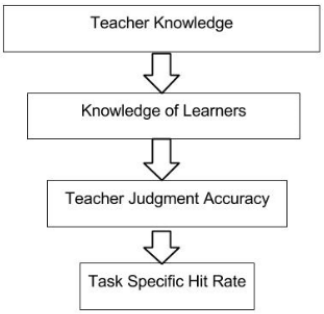


Figure 1 – Flow chart showing how the task specific hit rate measures the knowledge base of knowledge of learner.

Quality Science Teaching

A definitive model of effective science teaching has not yet been established by the science education community, however, the vast majority of published research in science education is consistent with the idea that social constructivism should guide teacher pedagogy (Ford, 2012; Hand, 2009; Kuhn, 2010b; Leach & Scott, 2002 McComas, 1998). As mentioned earlier, argument-based inquiry was chosen as a part of the theoretical framework because this teaching approach mimics the cognitive practices of actual scientists and has potential to improve the scientific literacy of students.

Previously discussed teacher roles in an argument-based inquiry classroom suggest that teachers take on a facilitative role to guide student inquiry. Teacher feedback is an instructional decision making practice that teachers utilize in this facilitative role that has been found to have significant impact on student learning, and may be an index of teachers' understanding of quality science teaching (Hattie & Timperley, 2007). This study examined teacher judgment accuracy as a measure of what teachers know about students, and feedback as an index of what they do with that knowledge. The instructional practice of dialogic feedback will be described in the next section.

Teacher Feedback

Shute (2008) defined feedback as "information communicated to the learner to modify his or her thinking or behavior to improve learning" (p. 82). Hattie (2009) claimed that many possible "agents" (e.g., teacher, peer, book, parent, self, experience) could provide the feedback information regarding aspects of one's performance or

understanding. Winne and Butler (1994) described the role of feedback in a classroom as follows: “feedback is information with which a learner can confirm, add to, overwrite, tune, or restructure information in memory” (p. 5740).

One way to measure a teacher’s ability to facilitate student knowledge construction is to evaluate the type of feedback given to the students. Research on teacher feedback in science classrooms that lead to greater student knowledge construction and performance on assessments suggests that a dialogic approach by the teacher is most appropriate (Chin, 2007; Hackling, Smith, & Murica 2010; Scott, Mortimer, & Aguiar, 2006).

Dialogic teaching, or feedback that follows a dialogic pattern, embraces the power of talk to arouse and extend students’ thinking and improve their learning and understanding (Alexander, 2006). It is logical to predict that teachers who use dialogic feedback in their classroom will have the ability to more precisely diagnose students’ needs and assess their progress (Alexander, 2006). In this way, teacher feedback should be related to a teacher’s knowledge of his or her own students.

The effects of teacher feedback, as a moderator of outcomes on assessment, have been studied extensively. According to Hattie (2009), feedback is one of the most influential instructional decision practices a teacher performs with respect to student learning and outcomes on assessment. In his large meta-analysis, Hattie (2009) claims that feedback had “effect sizes that were twice the size of other instructional influences” (p. 83). In another meta-analysis of teacher feedback, Murphy, Wilkinson, Soter,

Hennessey, and Alexander (2009) report that “dialogic discussion approaches produced strong increases in the amount of student talk and concomitant reductions in teacher talk, as well as substantial improvements in student comprehension” (p. 740). Thus, according to these meta-analyses, teacher feedback plays a large role in effective teaching.

A framework for the second predictor variable is offered below. In Figure 2, dialogic feedback flows out of instructional decision making, which is an aspect of argument-based inquiry pedagogy that is situated within social constructivist learning theory.

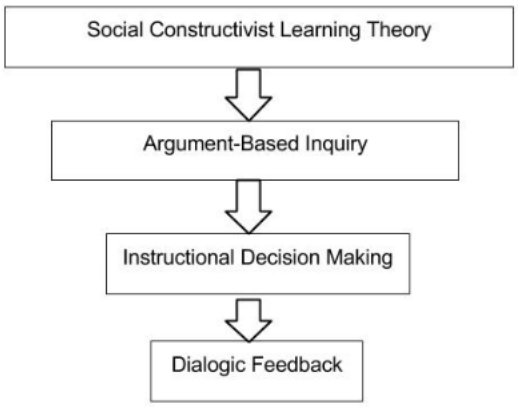


Figure 2 – Framework for the second predictor variable.

Rationale for Predictor Variables

Throughout the introduction a justification has been made for further study of teacher judgment accuracy and dialogic feedback patterns in science classrooms. The main hypothesis examined in this study was that a combination of knowledge of learners

(measured by teacher judgment accuracy) and instructional decisions based partly on that knowledge (dialogic feedback) would be related to student achievement.

The teacher judgment accuracy variable was intended to measure the depth of the teachers' knowledge of learners (KOL). Teachers who accurately predict how well their students will perform on assessments theoretically have gained access to their students' understanding of the concepts taught during the lesson. However, it is possible that a teacher could gain access to KOL through didactic lecture and repeated formative assessment in the form of similar multiple-choice quizzes. These instructional choices are not endorsed by the NRC (2012), the National Science Teachers Association (NSTA), nor the effective science practices endorsed by the NGSS (NGSS Lead States, 2013). Further, high judgment accuracy does not indicate the teacher is moderating his or her instruction based on the data or is addressing the individual needs of the students in the classroom. Evaluating each student by his or her level of understanding of the content assessed on each item will require the teachers to search their schema of their understanding of their knowledge of learners of each student. However, being able to predict student outcomes does not indicate whether the teacher is employing interventions to help struggling students; it is simply a measure of the level of understanding of the students' knowledge of the subjects taught in the unit.

The dialogic feedback measure is also a variable that alone, may not have predictive power. Teachers who score high on the dialogic feedback measure may demonstrate an ability to facilitate classroom discussion among peers, but this practice alone may not be enough to improve student understanding. Whereas the instructional

decision making practices aligned with dialogical teaching may improve student learning, there will always be students in a classroom who do not fully comprehend a lesson. Without knowledge of learners, a teacher may over- or underestimate students' understanding of the lesson and move on without addressing students' needs or misconceptions. Strength in knowledge of learners should allow a teacher to better tailor his or her instructional decisions to students.

The model for effective science teaching used in this study was one in which teachers gained access of knowledge of learners through various interactions. The teacher then tailored instruction based on students' understanding or misconceptions. Subsequent instruction was dialogic in nature (i.e., a student expresses a misconception and instead of correcting the student, the teacher asks the student to provide evidence of his or her claim) but did follow a step-by-step format. Instruction would instead change based on the teacher's knowledge of learners, which I predicted should lead to enhanced student outcomes. A framework for this interaction is shown in Figure 3 below.

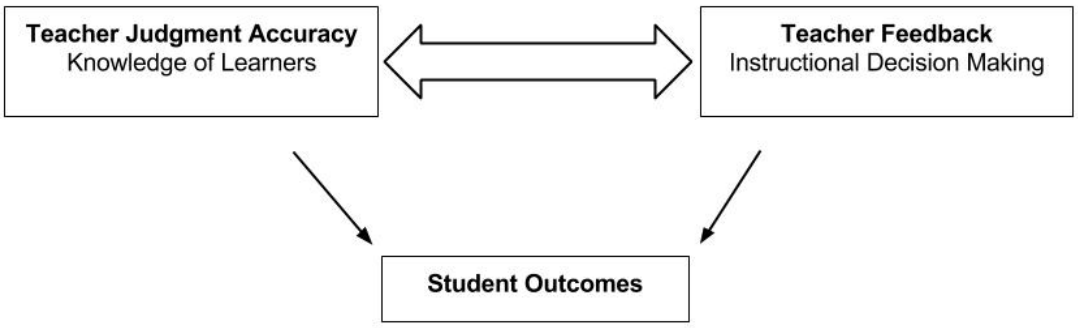


Figure 3 – Framework for teacher/student interactions and data collection

Student Assessment

The outcome variable in this study was student achievement on the science section of the Iowa Assessments. The Iowa Assessments were chosen because standardized tests are objective in nature. Standardized tests are often scored electronically by individuals who do not directly teach the student. They are also developed by experts and questions undergo a review process to remove bias.

Student achievement on standardized assessments is important for a number of reasons. First, if one of the goals of science education is to prepare students for careers in STEM fields, it is likely those students will need, minimally, a bachelor's degree from an institution of higher learning. Entrance into colleges and universities requires minimum scores on standardized assessment for incoming freshman and if students plan on pursing graduate degrees, many STEM fields require high scores on the Graduate Records Examination (GRE).

Second, while some areas of the country are reporting gains in science scores on national exams, overall U.S. students' scores are stagnant or declining. Improved scores on standardized testing across the country would reflect an improvement of scientific knowledge, thus working towards the second rationale for the focus on school science, which is to foster the growth of a scientifically literate population.

Research on the effects of inquiry teaching in science on student outcomes on standardized tests is limited, but initial findings are promising. Marx et al. (2004) report improvement in standardized science scores, knowledge of the content, understanding of

the process, and overall achievement of over 8,000 students, when an inquiry approach was implemented in a large inner-city school district. In a similar study, middle school teachers who used an inquiry approach increased the achievement scores of African American students, and narrowed the achievement gap between male and female students (Kahle et al., 2000).

In the studies cited above and others like them (e.g., Taylor et al., 2011) the moderating variable influencing student outcomes was the ambiguous term "inquiry." In the study, teachers were evaluated at a much more precise level, as I examined how their knowledge base and instructional decision making was related to student outcomes. A model displaying how the predictor variables flow from the literature as well as their predicted relationships is offered below in Figure 4.

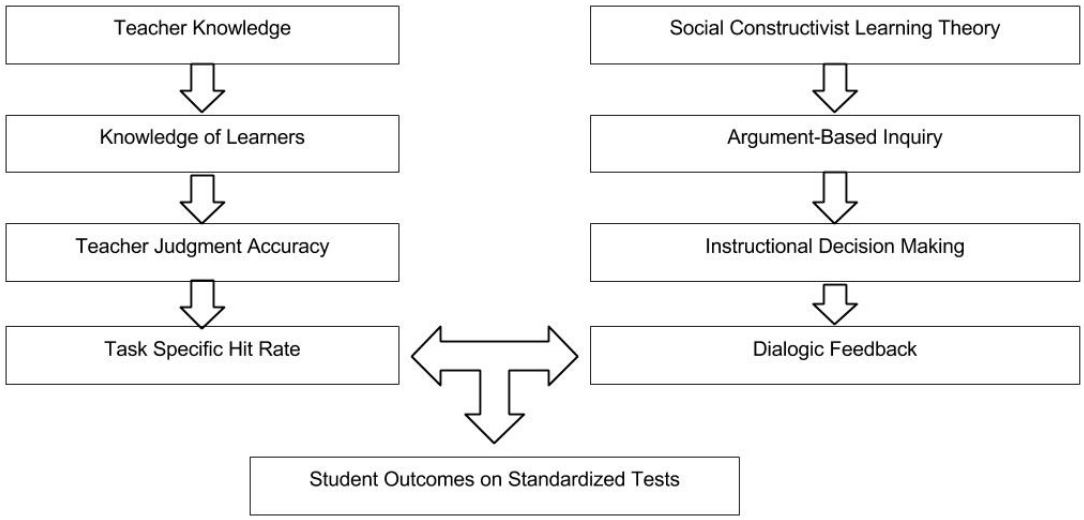


Figure 4- Model of two predictor variables and the outcome variable.

Statement of the Problem

Contemporary science education research has placed a premium on student-centered inquiry for decades. Most science education scholars agree that student ideas and questions should drive instruction and that students should construct their understanding by developing claims that are backed with evidence (Duschl et al., 2007; Ford, 2012). However, little is known about how accurate science teachers actually are at knowing their students' understanding of the concepts taught. Further, little is known about how teachers use this knowledge to adapt their teaching practices.

These problems motivated this study for two reasons. First, there are gaps in science education research regarding the teacher knowledge base of knowledge of learners. Teachers who value their students' ideas and use them to scaffold instruction should gain access to both their students' knowledge of the subject and the cognitive paths they took to arrive at their conclusion. Studies regarding these claims have not yet been published in the science education literature so the current study adds a missing piece to this literature and could lead to future studies evaluating how a teacher uses knowledge of learners to moderate instruction.

Second, this study evaluated the relationship between dialogic feedback and knowledge of learners on student outcomes. Previous research on dialogic teaching has claimed improved student outcomes (see Anderson, 2002). Most of these claims evaluated the teacher/student dialogue and have focused on how the teacher promotes discussion among students. This study went a step further and evaluated the level of

knowledge of learners, by measuring teachers' judgment accuracy, and the degree of dialogical teaching, by measuring teacher feedback, to determine if there is any predictive relationship among these two variables and student outcomes.

Research Questions

1. What amount of variance does teacher judgment accuracy and dialogic feedback predict, with respect to student outcomes on the science portion of the Iowa Assessments third through eighth grade?
2. Does judgment accuracy moderate the relationship between teacher feedback and student achievement?
3. Does dialogic teacher feedback mediate the relationship between judgment accuracy and student achievement?

CHAPTER 2

REVIEW OF LITERATURE

Judgment Accuracy

Teacher Judgment Accuracy (JA) has been described as an ability to judge student characteristics correctly and is part of a set of skills known as "diagnostic competence" (Artlet & Rausch, 2014). It has also been referred to as a teacher's ability to accurately predict task demands of their students (Anders et al., 2011). Teacher judgments of student achievement play a key role in many significant classroom decisions, including instructional preparation, differential grouping, recommendations, and retention (Eckert, Dunn, Codding, Begeny, & Kleinmann, 2006; Hoge & Butcher, 1984; Meisels, Bickel, Nicholson, Xue, & Atkins-Burnett, 2001). The ability of teachers to access an accurate representation of student achievement is widely understood to be essential for effective instruction, as it allows teachers to give necessary feedback and adapt their instruction to match student needs (Behrmann & Souvignier, 2013; Karing et al., 2013 Leinhardt, 1983).

Teachers' JA goes beyond accurately predicting students' outcomes on summative assessments; it may play a role in their understanding of students' comprehension of the lesson taught, which many consider a core aspect of adaptive teaching. Teaching in such a manner involves collecting information about students and adjusting lesson planning so individual needs of each student are met (Vogt & Rogalla, 2009). Published literature suggests adaptive teaching is an effective way to improve

outcomes on summative assessments (Brühwiler & Blatchford, 2011, Tobin & Llena, 2010, Parsons, 2012). Without accurate knowledge of their students' understanding of the content and concepts taught teachers may not be able to adapt their teaching practices to meet their students' needs. This information is important so lessons are taught at an appropriate level and students do not disengage because the lesson is too easy or difficult.

Given the importance that teacher JA plays in daily instructional decision making it is surprising that such little research has been conducted regarding this interaction. Current literature on teacher JA has identified two primary variables that affect the accuracy of teacher judgments, methodological variables and moderating variables (Hoge & Coladarci, 1989; Südkamp, Kaiser, & Möller, 2012).

Variables classified as "methodological" are categorized by the way the measure was collected. The studies in the two meta-analyses are methodologically alike because each contains a variable representing a teacher's judgment of a student's academic performance and examines that prediction of student performance on a standardized achievement test.

Moderating variables are external variables that impact the accuracy of teacher judgment of students' academic performance. These variables represent a process or a factor that alters the impact of an independent variable on a dependent variable. Typical moderator variables could include the sex of the teacher or student, age of the teacher or student, culture, or language proficiency of the students. The impact of both variables on JA will be discussed next.

Methodological Variables

Hoge and Coladarci (1989) categorized the methodological variables in the sixteen studies used in their meta-analysis as: direct versus indirect evaluations, specificity of the JA task, and norm-referenced versus peer independent judgment. Südkamp et al. (2012) reviewed 75 studies and identified similar methodological variables, but expanded to include two more. Their methodological variables were informed versus uninformed, number of points on a rating scale, instruction specificity, norm-referenced versus peer-independent, and domain specificity. Each of these methodological variables is described below.

Directness. The directness of the evaluation or judgment refers to the way in which teacher judgments were evaluated. Direct evaluations typically correlate teachers' judgment of a student's performance on a specific assessment with the student's actual performance on the assessment. For example, a teacher might be asked to predict students' scores on an upcoming math assessment; then the students' performance would be compared to the predictions. An indirect JA task would ask the teacher to make a prediction of students' performance that is not directly measuring their outcome on an actual assessment. The students would still complete the assessment, but the original prediction by the teacher would be more global. For example, correlating teacher judgment of students' general intellectual ability to students' actual score on a specific reading measure would be an indirect method. Another example would be a teacher grouping students into "high," "middle," or "low" groups and then comparing those groupings to student performance on an assessment. In each example the teacher was

asked to make a prediction about general ability, not student performance on a specific assessment.

According to Hoge and Coladarci's (1989) meta-analysis, in studies that used indirect judgments, the judgment/criterion correlations ranged from 0.28 to 0.86; the median correlation was 0.62. In contrast, the studies involving direct assessments generated a range of judgment/criterion correlations of 0.48 to 0.92, with a median correlation of 0.69. Although the medians are similar for direct and indirect judgments, there are far fewer low correlations for direct judgments, suggesting these tasks produce more accurate judgments.

Informed versus uninformed. Südkamp et al. (2012) used a slightly different categorization of directness in their meta-analysis claiming that: "The main difference between direct and indirect judgments is that teachers are either informed or uninformed about the test or the standard of comparison on which their judgment is based" (p. 745). The data Südkamp et al. (2012) gathered indicated higher correlations between students' academic achievement and informed teacher judgments (mean effect size = .76) than uninformed teacher judgments (.61). The vast majority (88%) of the studies in the Südkamp et al. (2012) meta-analysis identified uninformed JA tasks as a methodological variable. This larger sample size could be a reason why the correlation scores are lower. In both meta-analyses the median scores for direct or informed JA tasks was higher than indirect or uniformed suggesting that teachers can predict their students' outcomes more accurately if they know the task criterion.

Specificity. Judgment specificity refers to the explicitness with which teachers provide their estimate of students' academic achievement. Both meta-analyses categorized JA tasks into categories based on their degree of specificity. Studies in the Hoge and Coladarci (1989) meta-analysis that utilized tasks ranging in specificity revealed a wide range in scores. The median *r* correlation coefficent fell between .28 and .92 (see Table 1). Hoge and Coladarci (1989) found a high median correlation for all levels of specificity (see Table 1), however it should be noted that their study only looked at 16 published articles at the time, with a disproportional number using the lowest level of specificity.

Similar to Hoge and Coladarci (1989), Südkamp et al. (2012) used the same categories of specificity in their meta-analysis. Overall, the mean score for all levels of teacher judgment specificity was high in both meta-analyses (Hoge & Coladarci [1989] *r* = .61, Südkamp et al. [2012] *r* = .63). A more detailed description of JA task specificity will be discussed below.

JA specificity tasks have been labeled in the ascending degree of specificity as: rating, ranking, grade equivalences, number correct on a given measure, and item responses or hit rate. As the specificity of the JA task increases the teacher's level of detailed knowledge about the student must also increase (Artelt & Rausch, 2014). JA tasks of low specificity (ratings or rankings) require less detailed knowledge of student capability. These types of tasks may ask teachers to rate students into global categories (e.g. low, middle, or high), or rank students in comparison to their peers. Mid-level specificity tasks (grade equivalency) require the teacher to activate a more detailed

prediction of their students' academic ability. Teachers would need to use information that students have provided in past interactions (i.e. assessments, informal discussions, journal writing) specific to the task to predict how they will score on future assessments. JA tasks of high specificity require the most detailed level of teacher cognizance. The fourth-most specific JA task is the number of correct responses assessment. In this task teachers would make direct judgments of their students' performance on an assessment by predicting the number of items the students will answer correctly. The most specific JA task will be discussed below.

The item-specific hit-rate requires a teacher to predict the actual items a student would answer correctly on an assessment. Hit-rate tasks are interesting because of the level of understanding about a student that a teacher would need to accurately make the prediction. Karing et al. (2013) noted that hit rate tasks use "an indicator of all the available information of the judgment of individual student's performance on individual tasks" (p. 74). Thus, the task-specific hit rate takes into account whether the tasks that were evaluated by the teachers are the same tasks that were performed by the students; giving a more detailed report of the teacher's understanding of specific skills the student has mastered (Artelt & Rausch, 2014). Teachers who have the ability to make predictions at hit rate specificity would need students to first make their understanding of the content available to them and then they would need to utilize this information in a way that was beneficial to the learner. An example of this would be a teacher asking an open-ended question to elicit information about student understanding, then utilizing that information

to construct feedback that prompts a deeper, more thoughtful response from the student to determine if the student actually has conceptual understanding of task.

The two upper-level specificity tasks (number of correct responses and hit-rate tasks) require more acute detection skills than the three JA tasks of lower specificity. Although the two upper-level tasks differentiate themselves from mid and low-level specificity tasks in terms of detailed knowledge the teacher must have of the student there is a small, yet possibly significant, difference between the two. On a number of correct responses task a teacher may accurately predict that a student will correctly answer seven out of ten questions correctly, but if the student misses three different questions than the teacher predicted the level of knowledge that the teacher has of the student's understanding may not be as accurate as the score indicates. If the questions on the assessment are heterogeneous a mismatch of correct responses may further confound the results. The task-specific hit rate eliminates this possible variance because predictions are item specific.

Literature focusing on hit rate tasks is very limited. In the two studies Hoge and Coladarci (1989) evaluated, Coladarci's study (1986) only had eight teachers (five third grade and three fifth grade) and Leinhardt's study (1983) involved just 11 teachers and 11 novices (non-teachers or pre-service). In the Südkamp et al. (2012) meta-analysis, studies using hit-rate were not included, presumably because of the lack of studies that utilized this variable. According to Südkamp et al. (2012) the group of the lowest specificity, ratings, is most commonly utilized in JA studies (86.8% of the studies in their meta-analysis used ratings). It is not clear why the majority of the JA studies have used

global ratings instead of more specific indicators of student knowledge, but it is possible that the outcomes of the studies did not require a specific judgment.

The lack of empirical studies of teacher diagnostic sensitivity at the task-specific hit-rate level was a central motivation of this study. The teachers' ability to make more specific judgments may require a deeper understanding of student comprehension. Learning how teachers achieve diagnostic competence at the highest level of specificity was an important factor of teacher knowledge to study, because it may be possible that teachers who understand students' competence at an item-specific level may be more competent at utilizing adaptive teaching practices.

Table 1- *Results from Hoge and Coladarci's (1989) meta-analysis, judgment specificity domain.*

Specificity	Number of studies	*r* score
Ratings	8	.61
Rank order	2	.76
Grade equivalents	3	.70
Number of correct responses	1	.67
Item Based Judgments	2	.70

Norm-referenced vs. peer-independent judgments. In addition, teacher judgments may differ in whether they are norm-referenced or peer-independent. Norm-referenced

judgments require teachers to make a judgment of each student's ability in comparison to his/her peers, or a norm group, whereas peer independent judgments required teachers to predict their students' achievement without a comparison group. Hoge and Coladarci (1989) evaluated this variable in their meta-analysis but found no substantial difference between correlations. The median correlation for norm-referenced judgments was .68 and peer-independent judgments was .64. Norm-referenced judgments typically would be tasks of low specificity since the teacher is required to compare a student to his or her peers. Low specificity does not necessarily mean that the teacher is less aware of student ability. For instance, if a teacher scored high on a norm-referenced rank order judgment of their students' ability in division fluency it would indicate they know which students are "high," "medium," and "low" and hopefully the teacher could adapt future lessons appropriately. However, if a teacher scored high on a similar task of the students' general ability in "math," it would not indicate specific areas where low students need improvement. Peer-independent judgments would almost always be tasks of middle to high specificity and would require teachers to predict specific outcomes of each individual student. Südkamp et al. (2012) noted that peer-independent teacher tasks may "lead to higher correlations between teacher judgments and students' academic achievement, because this approach allows teachers to focus on each student individually, preventing judgment biases due to the achievement of other students in the class" (p. 745).

In the Hoge and Coladarci (1989) meta-analysis, for studies using peer-independent ratings, the median score was $r = 0.68$, with a range from $r = 0.67$ to 0.72;

for norm-referenced judgments, the median correlation was $r = 0.64$, with a range from $r = 0.28$ to 0.92. Similar to the Hoge and Coladarci (1989) meta-analysis, Südkamp et al. (2012) found no statistical significance for the methodological variable of peer-independent versus norm-reference JA tasks. In this study, peer-independence would not yield enough details about the type of judgment. Peer-independent judgments require the teacher to focus on an individual student, but the specificity of the judgment could vary from global to item-specific.

Points on a rating scale. Südkamp et al. (2012) also evaluated the global ranking of rating scales. The researchers found a large disparity in the number of points on a rating scale ranging from 2 – 100. Ratings with a large number of categories allowed for a more detailed judgment while those with a low number of points offered a more global judgment. The effect of the number of points on the rating scale was not statistically significant.

Domain specificity. The final methodological variable used in the (Südkamp et al., 2012) meta-analysis was the precision of the teacher judgment as an inclusive academic ability or specific domain (i.e., mathematics or language arts). Most studies that assessed a domain specific JA task evaluated language arts/reading or mathematics. Only a few studies have looked at teacher JA in science and social studies (i.e., Hopkins, George, & Williams, 1985; Wright & Wiese 1988) and those studies included aggregated students' math and reading scores. Thus, no published studies could be located that looked at teacher JA in the domains of science or social studies. These findings alone raise questions about how teacher JA affects instructional decision making in science

classrooms. Science is a discipline that many believe is best taught through an inquiry approach. Any effective inquiry-based science teaching approach focuses heavily on the individual student's conceptual understanding of the lesson (Berland & Reiser, 2009; Hand, 2007; Hand, 2009; Kuhn, 2010a; Kuhn, 2005; NRC, 2012), suggesting that teacher JA may be a potential critical component of effective science teaching.

Accurate scores on JA tasks of global academic ability provide teachers a general sense of their students' academic competence, but do not provide the teacher any useful knowledge of their domain specific ability. A student may have advanced mathematical skills, but have poor language arts capability. A global judgment of this student would not help address his or her shortcomings or possible advanced placement. Even a global judgment of overall ability in a specific domain provides limited information for the teacher because it is unlikely that the student exceeds expectations in all areas of that domain. In the (Südkamp et al., 2012) meta-analysis, 27.4% of the studies were based on judgments of overall academic ability, 23.2% on judgments of an academic ability in one subject, and 49.5% on judgments of a specific academic ability within a subject. None of the studies in the meta-analysis were carried out in the domain of science. This leaves many questions unanswered about what teachers know about their students' understanding of their science lessons and how accurate teachers are at predicting their students' ability to perform on a standardized science test.

The reason the domain of science was investigated in this study was based on the epistemological belief of many science education researchers that science is a discipline rooted in argumentation. Students engaging in argumentation with their peers has been a

focus of much scholarship in science education (e.g. Berland & Reiser, 2009; Hand, 2009; Kuhn, 2010b). Many researchers believe that science is a social process that constructs knowledge in and through people (Kuhn & Crowell, 2011; Newton, Driver, & Osborne, 1999). It is through language that students may come to understand science. Language is essential for scientists to explain and argue for their ideas. Real scientists use language not only for inquiry but also as an inquiry approach (Wallace, Hand, & Prain, 2004). Language is a critical component of the ways in which one becomes literate in science (Jiménez-Aleixandre & Erduran, 2011). A teacher who allows more frequent peer-to-peer science argumentation opportunities in the classroom may have access to information that reveals a student's true conceptual understanding of the topic. Analyzing the effects of teacher feedback on the specificity of teacher JA would be a first-step in understanding how effective science teachers obtain information from their students that they deem useful for future instruction.

Summary of Methodological Variables

In both meta-analyses only one methodological variable moderated teacher JA at a statistically significant level: direct versus indirect (or informed vs. uninformed) judgments. As noted earlier, Südkamp et al. (2012) distinguished between the two terms because informed versus uninformed studies are confounded by judgment specificity. In their meta-analysis, they found that considerably more studies used uninformed judgments than informed judgments. Also mentioned earlier, Hoge and Coladarci's (1989) results were not as profound as those of Südkamp et al. (2012), but in both meta-analyses teachers were more accurate at predicting direct or informed judgments.

Even though the sample size is small, the data from the meta-analyses suggests that teachers are more accurate at predicting their students' outcomes when they know the specific criterion that they will evaluate. Teachers' ability to make more specific predictions on informed/direct JA tasks may mean they have obtained detailed knowledge about their students' understanding of the task, but until the variable of specificity is evaluated more closely the claim does not warrant much merit. Further, the purpose of this study is not to investigate whether teachers are more accurate predictors of student achievement. Instead the purpose of this study is to determine whether teacher JA plays a role in their instructional decision making and how these variables are affected by teachers' feedback patterns. If teachers are better at predicting direct judgment tasks then it may mean they have obtained specific information about their students through their instructional decision making practices. Questions still remain about how the teachers made these decisions and what influenced them.

Moderator Variables

Several studies have found teacher experience (e.g., Leinhardt, 1983), academic subject area (Demaray & Elliott, 1998; Dohert & Conoll, 1985; Eckert et al., 2006; Sharpley & Edgar, 1986; Wright & Wiese, 1988), sex of students (Dohert & Conoll, 1985; Hoge & Butcher, 1984; Sharpley & Edgar, 1986), and student achievement level (Coladarci, 1986; Demaray & Elliott,1998; Feinberg, & Shapiro, 2003; Leinhardt, 1983) have no statistical significance and impact the accuracy of teacher judgment. These data are promising because they suggest that outside influences do not seem to sway a teacher's ability to predict students' academic capability. There have, however, been a

few studies that looked at the moderator variable of teacher instructional decision making practices and how it affects student outcomes.

JA, Moderator and Methodological Variables, and Student Outcomes

The majority of studies mentioned previously in the two meta-analyses indicate that teachers' judgments correlate fairly highly with the academic performance of their students: in their 1989 meta-analysis, Hoge and Coladarci reported correlations between *r* = .28 and r = .92,with a median correlation of *r* = .66. In the Südkamp et al. (2012) study, the Fisher's z-transformed correlations ranged between *r* = -.03 and *r* = 1.18, with a median correlation of *r* = .53. Although the median correlation score for both studies is fairly high, two points should be made. First, the vast majority of studies in both meta-analyses were studies in which the teachers were asked to complete tasks of low judgment specificity and the extreme range between scores suggests teachers are making decisions in their classroom that allow them either more or less access to their students' content knowledge.

Currently, there have been only a few studies that have examined the relationship between these variables (see Table 2). Karing et al. (2013) looked at the relationship between task-specific hit rate, how individualized teachers presented their lessons, and student outcomes, measured by a short scale consisting of four items. The Likert-type response scale ranged from 1 ("I disagree") to 4 ("I agree"). The teachers also reported how they used structuring cues in their lesson by assessing three items on a Likert-like scale and self-developed items (an example item is: "I summarize the lesson so they can

remember the gist"). The Likert-type response scale ranged from 1 ("never") to 4 ("very much"). The teachers were also asked to complete two JA tasks with varying methodological variables (rank order and task specific hit rate). Karing et al. (2013) reported German language teachers showed a mean task-specific hit rate of $M = 0.66$ ($SD = 0.11$), implicating that they could accurately judge 66% of their students' answers in the domain of reading. Task specific hit rate had a significant positive relation with the development of students' reading literacy. For teachers who used a high degree of individualization during lessons, Karing et al. (2013) found a "Significant positive relation between the task-specific hit rate and the development of students' reading literacy, but teachers who applied a low degree of individualization, the task-specific hit rate was not significantly related to the development of students' reading literacy" (p. 279). Karing et al. (2013) reported significant positive relations between rank-order JA and students' reading achievement. They did not observe this in the group with a low degree of individualization or in the group with a high degree of individualization. The data in this study indicated the group of students that showed the highest achievement in literacy had teachers who displayed high JA on tasks of the greatest specificity (hit-rate) and taught lessons that were very individualized. Rank order and individualization did not seem to make a significant difference in the students' achievement in literacy. It could be interpreted that these teachers obtained very specific information about each student and utilized that information in a way that was beneficial to each student.

Behrmann and Souvignier (2013) looked at how teacher feedbacks and task specific hit rate affected student reading outcomes. The study found that a high number

of feedbacks and high JA produced significant gains in student achievement, low feedbacks and high JA produced much lower achievement gains, and high feedbacks and low JA produced no growth at all.

In a similar study, Helmke and Schrade (1987) evaluated the correlation of teacher structuring cues and diagnostic sensitivity at the hit rate level and determined whether it affected student growth on standardized assessments. Cognitive growth was highest when high JA was combined with high frequency of structuring cues. When JA was low, high or low frequency of structuring cues didn't make a difference. Neither diagnostic sensitivity nor the mere frequency of structuring cues appeared to affect cognitive growth substantially. High achievement gains were associated only with a combination of both. The group that performed the lowest on student outcomes was the group that had high JA and low structuring cues. The final finding was somewhat puzzling because studies have shown that students of teachers that display high JA typically score high on academic assessments (Coladarci,1986; Demaray & Elliott, 1998; Karing et al., 2013). It should be noted that the moderator variable in the Helmke and Schrader (1987) study was measured by simply quantifying the number of structuring cues. There was no measure of the quality of the interaction between the teacher and student (see Table 2). Helmke and Schrader (1987) mentioned it is possible that if the teacher has high JA and does not do anything with the information (i.e. observes that a student doesn't understand the information, but just continues with the instruction and doesn't provide feedback to the student) it could cause lack of motivation and thus poor performance on assessments.

Table 2 - *Summary of the findings in the three studies that evaluated JA moderator and methodological variables and student outcomes.*

Study	Domain	Moderating Variable	Specificity	Outcome
Karing et al., (2013)	Literacy	High Individualization of lesson	High Hit Rate	*Significant Gains
	Literacy	Low Individualization of lesson	High Hit Rate	No significant gains
	Literacy	High Individualization	High Rank Order	No significant gains
	Literacy	High Individualization	High Rank Order	No significant gains
	Literacy	Low structuring cues	High Hit rate	*Significant Gains
	Literacy	Low structuring cues	Low hit rate	No significant gains
	Literacy	High structuring cues	High Hit Rate	No significant Gains
	Literacy	High Structuring cues	High Hit rate	No significant gains
	Literacy	High and Low structuring cues	High and low rank order	No significant gains
Behrmann & Souvignier (2013)	Reading	High feedback	High Hit Rate	*Significant Gains
	Reading	Low Feedback	High Hit Rate	Lower gains, not statistically significant
	Reading	High Feedback	Low Hit Rate	No significant gains
Helmke & Schrader (1987)	Math	High Structuring Cues	High Hit Rate	*Significant Gains
	Math	High Structuring Cues	Low Hit Rate	No significant gains
	Math	Low Structuring Cues	High Hit Rate	no significant gains
	Math	Low structuring cues	Low Hit rate	No significant gains

Summary and Potential Design of JA Interactions

Teacher judgments of student achievement may have a considerable impact on students' learning experiences and academic potential. Many professional decisions are based on teacher judgments of student characteristics and knowledge; for example, decisions related to ability grouping, adaptive teaching, and grade allocation (Eckert et al., 2006).

Published literature on teacher JA has reported that teachers are fairly good predictors of student outcomes on tasks of low specificity. Some researchers (i.e., Llosa, 2008) have argued that teachers may be accurate collective predictors of student achievement but not at the level of specific content areas or standards. The literature on teachers' ability to make item level predictions is so scarce it is difficult to make any claims about how this ability is developed or how accurate teachers really are at this level of specificity.

In the two meta-analyses of teacher JA the only methodological variable that revealed any statistical significance of variance was direct/indirect or informed/uninformed. It would be logical to assume that tasks that were uninformed or indirect would be tasks of low specificity, like ratings or rank order. Conversely, it would be nearly impossible to ask a teacher to make a prediction on tasks of high specificity (number of correct responses or item-level judgments) and keep the teacher unaware of the students they were judging. The statistical discrepancy of direct/indirect –

informed/uninformed judgments found in the two meta-analyses may be influenced by the level of specificity of the JA task.

As a practicing teacher and future researcher what is most interesting is examining how teachers obtain information from their students, what they do with that information, and if certain instructional decision making practices allow the teachers access to a greater understanding of their students' knowledge of the topic.

A potential model for how a teacher obtains information about his or her students is offered by Funder's (1995) Realistic Accuracy Model (RAM). Funder (2012) described the RAM model as a way to "achieve accurate judgments based on behavioral information detected by a judge, who utilizes that information correctly" (p. 177). He went on to explain that the RAM model "describes the process that connects a personality trait of a person with a correct judgment of that trait in the mind of a perceiver" (p. 179).

The four conditions Funder (1995) described are: relevance, availability, detection, and utilization. The RAM model provides a framework for understanding how the teacher/student interaction could present itself in a classroom and how the teacher could acquire knowledge of students' understanding of the lesson. The first three conditions, and their relation to teacher JA, will be discussed briefly and a more in-depth description of the fourth condition, utilization, and how that condition includes aspects of the previous three will be discussed. A short description of how teachers' instructional decision-making practices may affect the four conditions of the RAM model will also be included.

Relevance. Artlet and Raush (2014) described relevance as the first condition for accurate teacher judgments. They claim “In order to be able to arrive at an accurate judgment of a student’s characteristics, the student must reveal some kind of information that is informative about the respective characteristic” (p. 33). The student must reveal something that is relevant to the objective of the lesson. If a teacher asks a question about force and the student provides an explanation about energy the revealed information would not be useful to the goal of capturing the designed outcome. Students can reveal information in a variety of ways (i.e., quiz, test, writing, dialogue); this information is critical for the teacher to make an accurate judgment of their understanding.

Availability. In order for the teacher to access information about student understanding the information must be available (Artlet & Raush, 2014). If the information is not available to the teacher it will be impossible for the teacher to form an accurate judgment of the students’ understanding and to adapt future lessons that fit the students’ needs. The exchange must happen in a context that the teacher shares with the student in order for the information to be available. For example, if students feel comfortable explaining their understanding with peers in small groups, but not in front of the entire class the teacher may miss the more private interactions and inaccurately judge the student’s understanding of the lesson.

Detection. In the third condition of the RAM model, teachers must detect what information is important for future teaching. Examining the relationship of a teacher’s score on a JA task of high specificity and the teacher/student interactions may reveal something about their ability to detect relevant information about the student. Similar to

availability the student and teacher must share a context in order for the teacher to accurately detect the student's understanding. If the primary focus of a teacher's lesson is a lecture it would be unlikely that he or she would have keen detection skills of the students' comprehension. Once the information is detected by the teacher the next step is executing something with that information.

Utilization. The fourth, and most relevant condition for the proposed studies is utilization. A key aspect of the teacher/student relationship is what the teacher does with the information obtained from the student. "Teachers have to correctly utilize the relevant, available, and detected information, and interpret it accurately, in terms of what it implies about the child's competence" (Karing et al., 2013, p. 35). The more detailed information a teacher has about his or her students the more data he or she will have to make decisions about future instructional choices. A teacher could score high on a global JA task but not be able to detect specific areas of weakness, recognize acquired skills and knowledge that the student has that the teacher could build upon, or utilize information gained from the student in a meaningful way. A teacher with high JA on a task of high specificity may have the information needed to adapt future lessons, but simply gaining access to the information does not mean the teacher necessarily utilizes it in a way that benefits the student.

An example of positive utilization might start with asking students to write in a journal to express their ideas about how light allows humans to see objects (relevance); the students then share their ideas with the teacher through dialogue or their writing (availability); and the teacher now has a clear picture of the student's pre-teaching beliefs

(detection). Next, the teacher may ask the students to collect evidence to support their ideas through experiments or reading the science book (utilization). In this way the teacher has used previously gathered information about students' understanding to inform his or her instructional decision making. A model for these interactions was created by Funder (1995) and is shown in Figure 3 below. In the next section a description of how a teacher's instructional decision-making practices could affect each component of the RAM model will be discussed.

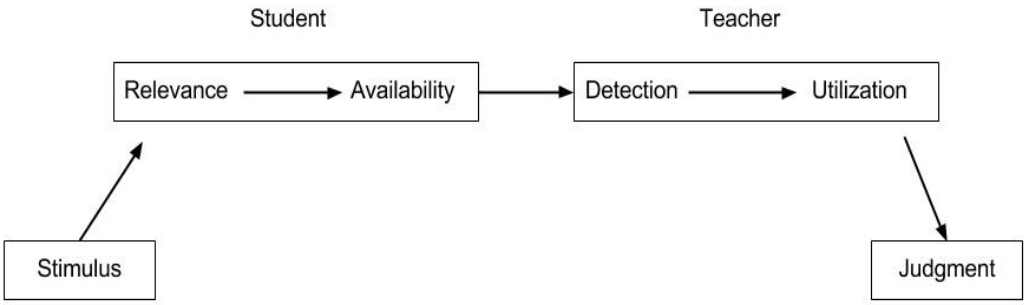


Figure 5 – Model of RAM interactions (Funder, 1995, p. 659)

Teacher Instructional-Decision Making Practices

The decision-making practices that teachers make will have a major impact on how students initially reveal their understanding of the lesson. A teacher who delivers didactic teaching and only asks students to demonstrate their understanding through

standardized quizzes and tests would ask students to reveal different information than a teacher who asks students to indicate their understanding through informal writing, small group dialogue, or classroom experiments. It is possible that a teacher's instructional decision making practices may impact their ability to accurately predict their students' outcomes on JA tasks of varied specificity. For example the former teacher might gain access to specific information if an item analysis of the assessment was done. The latter may ask students to reveal different information through a variety of techniques that are more personal to the learner and may illustrate different information. The pedagogical choices that the teacher makes will likely result in different ways students present their understanding. These pedagogical choices may also moderate the way a teacher proceeds with instruction.

The instructional decision making practices of a teacher also have an impact on how available the information is to the teacher. For instance if a teacher only makes information about student understanding available through quizzes and tests there is a possibility that the student may misinterpret the instructions of the test question (i.e., failure to notices negatives in the questions) or have test anxiety. The students' actual understanding of the material may not be available to the teacher causing the teacher to make inaccurate judgments about the student. If the teacher does not have accurate information available to them it will be difficult to enact adaptive teaching and the JA ability of the teacher will likely be lower. It is possible that teachers who use practices that allow for greater variety of access to student understanding (i.e., allowing students to raise questions and negotiate their understanding with peers or teachers having longer

dialogical interactions with students by asking a series of probing questions) may make students' understanding more available to them. In the RAM model, relevance and availability are primarily responsibilities of the student. The teacher can influence the ways that students reveal and make available their understanding of the lesson, but ultimately it is the student's job to present their knowledge to the teacher. The final two steps of the RAM model involve what the teacher does with the information obtained in the first two steps.

Once the student makes their understanding available it is the teacher's responsibility to detect the relevant information made available to them. A teacher who uses lecture and makes student understanding available via quizzes and tests may score high on JA tasks of a global level of specificity, but if they do not detect the area in which an individual student is struggling then their ability to help that student will be limited. For example, if that same student scored 6 out of 10 on a math quiz on subtraction and the teacher simply re-taught the lesson using the same techniques they might have missed the fact that the student missed four problems where they had to borrow across a zero. If the teacher had detected the mistake the re-taught lesson could have a much bigger impact on the student. It is possible that teachers who score high on JA tasks of high specificity may be better detectors of their students' understanding. In contrast, if a student demonstrated that they understood certain aspects of the lesson a teacher could use that information and build upon it to help them further understand the goals of the unit.

The ability to simply detect accurate student understanding of the lesson would likely not result in student success on summative assessment. In the next step the teacher would need to "do" something with this information. Scoring high on a hit-rate JA task would provide empirical evidence that the teacher has specific knowledge of his or her students' understanding. Depending on the task it may also provide evidence that the student has demonstrated content and conceptual understanding of a topic. For example, depending on the construct of the task a science teacher who asks students to explain why a dropped tennis ball does not bounce back to its original dropping point may or may not be able to detect if the student understands the concept of energy transfer. If the student can explain that energy from the moving tennis ball was transferred to sound and heat energy when it struck the floor it would demonstrate that they understand energy transfer in this single phenomenological example. If the student was only asked to reveal and make available their understanding of this single observation it would not provide the teacher with their students' conceptual understanding of energy transfer. But if the teacher asked the student to provide another example of this concept it would give the teacher a better picture of the student's conceptual understanding, thus making them better predictors of the students' success on future assessments.

The feedback that the teacher provides to the student has the potential to moderate the type of information that student reveals and makes available for the teacher to detect and utilize. There may be certain instructional decision making practices that allow teachers greater ability to detect student understanding and teacher JA ability may play a role in this.

The level of specificity that a teacher can predict his or her students' understanding and the decisions that they make with that information is an area that little research has been conducted and was motivation for this study. Three areas of interest in this study are the methodological variable of specificity, moderator variable of teacher feedbacks with students and the domain of science.

Knowledge of student understanding of the content is critical for teachers to predict student achievement with any accuracy. However, as found in the literature review, student achievement on assessments was not optimal with only high teacher JA. Students achieved success when their teachers had high JA and another moderator variable. In this study the moderator variable that was evaluated was teacher feedback. In the next section teacher feedback will be discussed.

Teacher Feedback

This section will examine the role teacher feedback plays in enhancing student discourse, improving scientific reasoning, and shifting student perceptions of scientific phenomena from everyday views and everyday reasoning to scientific views and scientific reasoning. The section will review published literature of the effectiveness of teacher feedback in science classrooms and examine how teacher feedback can situate the epistemological framework of the classroom into a "Vygotskian" social-constructivist environment where students construct their learning through negotiation. feedback definitions will be introduced as situated in science classrooms framed by inquiry, discuss Vygotsky's Social-Constructivist Learning Theory and Zone of Proximal

Development (ZPD) and how they are related to teacher feedback, investigate different levels of feedback evaluated in numerous meta-analyses and their impact on students' cognitive growth, look at patterns of teacher feedback (specifically dialogical and authoritative), and discuss how a science teacher deals with the delicate balance of allowing students' everyday knowledge and reasoning to guide instruction and still introduce the science content agreed upon by the scientific community. The section will begin with a working definition of feedback and a theoretical framework for teacher feedback in science education based on Vygotsky's Social Constructivist Theory.

Feedback

Shute (2008) defined feedback as "information communicated to the learner to modify his or her thinking or behavior to improve learning" (p. 156). Hattie (2009) claimed that many possible "agents" (e.g., teacher, peer, book, parent, self, experience) could provide the feedback information regarding aspects of one's performance or understanding (p. 81). Winne and Butler (1994) described the role of feedback in a classroom as, "feedback is information with which a learner can confirm, add to, overwrite, tune, or restructure information in memory" (pp. 5). These definitions and characteristics of feedback are not specific to science education, but contain aspects that many science education researchers would deem effective in promoting effective learning in a science classroom. Specifically, Winnie and Butler's (1994) view that feedback can be information that is tentative and is open to interpretation is a pattern of thinking that scientists use when presented with new information. This view of feedback in a science classroom will be discussed below.

Many reformed-based movements in science education have called for students to engage in classroom practices that align with practices of actual scientists (Ford, 2012; NGSS Lead States, 2013; NRC, 2012). This idea of making classroom practices more authentic is not new and can be traced back to Dewey's vision of "psychologizing" the curriculum (Shulman & Quinlan, 1996). Dewey believed that school curricula would combine teachers instructing students to learn the intellectual activities of experts in the discipline's field, along with aspects of how authentic learning occurs within the discipline, while meeting the needs of students (Shulman & Quinlan, 1996). Some have debated the merits of having students act as scientists or if students can be considered members of the scientific community (Ford, 2012), but there is little debate in science education that all students can engage in the practices of scientific reasoning and scientific negotiation if given the opportunity (Bricker & Bell, 2008; Duschl et al., 2007). If science teachers want to emulate how authentic discourse and feedback occur in the scientific community this pattern of communication warrants further discussion and will be discussed in the next section.

Feedback in the Scientific Community

The scientific community, like all communities of practice, exists to build upon and improve the current understandings and beliefs of the community. In order to move forward the community must decide what the aim of the practice is and what will be considered knowledge (Ford, 2010). When new claims challenge the beliefs of the community these ideas are met with the scrutiny of peer-review. A "new" idea cannot be considered until it can be proven to be more than a single occurrence or a

phenomenological event that does not fit into a model or pattern that has been accepted into law. Kuhn (1970) called this process “normal science,” and the behavior of the scientific community has followed this social pattern for centuries. A key aspect of normal science is that the discourse yields no single voice authority to govern over the community of scientists (Kuhn, 1970). Feedback between scientists is rarely authoritative and typically throughout the peer-review process scientists are asked for evidence to back their claims, scrutinized about their data collection process, or asked if their overall ideas fit the scope of practice of the discipline (Lemke, 1990).

Therefore, the practice of science, although complex in the specific details, is actually simple in its basic structure. Scientists attempt to explain accounts of nature and negotiate with other scientists about these accounts, drawing on information they have gathered on nature's behavior (Ford, 2012; Gross, 1990). After the community of peers comes to a consensus of the negotiated account, the original claim either becomes a part of the community’s knowledge or is rejected (Kuhn, 1970). The practice of the scientific community can be used by teachers as a learning tool in the classroom. These practices and how teachers can use feedback to help students engage in similar practices of the scientific community will be discussed in the next section.

Teacher Feedback Practices That Promote Student Dialogue

The previous summary of how actual scientific discourse occurs provided an example of how science is not a static collection of knowledge but is rather a dynamic and complex result of many dimensions of social negotiation. Yet, many teachers enact

curriculum that asks students to only consider one view of scientific phenomena. It would be easy for teachers to argue that the views of the scientific community are settled on most topics (i.e., gravity or energy) so training students to "fit into" a static social position is appropriate as long as they receive the correct information. The counter argument to this claim is that actual scientists never accept new information as passive receivers and always consider their beliefs as tentative (Ford, 2008; Ford, 2012; Hand, 2009; Hand, 2007). Packer (2001) argued that schools need to "recognize their responsibility for preparing students to transform themselves instead of focusing exclusively on the transfer of isolated skills, concepts, or strategies across static contexts" (p. 74). Therefore it is reasonable to conclude that if the goal of a classroom science teacher is to prepare students to become scientists or to think "scientifically" using a pedagogical approach that promotes convergent thinking or a static view of science is not the most effective choice. Some of the feedback patterns that follow this approach will be discussed in the following section.

Shute (2008) classified teacher feedback patterns that matched the previously mentioned divergent views as verification and elaboration. Verification feedback patterns simply state whether the claim is "correct" or "incorrect." Teachers who use verification feedback typically provide the correct answer to students when they give an incorrect answer, and the teacher attempts to focus the class on one point of view (Shute, 2008). Elaboration feedback provides students information in the form of cues or prompts that guide the learner toward a better understanding of the topic (Shute, 2008). Teachers who use patterns of elaboration feedback would look for, and build upon the knowledge that

the student already has instead of looking for knowledge that the student is missing. Shute's elaboration pattern of teacher feedback would likely be a better fit for teachers who are attempting to mimic the practices of the scientific community. Published research has maintained that a more facilitated elaborative teacher approach to feedback has obtained more positive results on student outcomes than a verification approach (Bangert-Drowns, Kulik, Kulik, & Morgan, 1991; Black & Wiliam, 2010; Hattie & Timperly, 2007; Kluger & DeNisi, 1996; Shute, 2008; Kluger & DeNisi, 1998). This review of literature on effective feedback in science education has led to the working definition that will be used in this paper.

Working Definition of Feedback.

For the purpose of this study feedback will be considered as: Information shared between teachers-students that helps guide the further construction of scientific ideas and the scientific process. Most information that is shared between teachers and students occurs through dialogue, however ideas can also be communicated through writing, drawing, graphs, or other modes of communication, so the term "information" was chosen as a more inclusive term than "language."

In the next section a theoretical framework based on Vygotsky's Social Constructivist Theory will be provided. Vygotsky's theory was chosen because it is consistent with the previously mentioned patterns of social discourse in which practicing scientists engage. Most science education researchers believe these patterns of scientific

dialogue should be emulated in the classroom so students have the opportunity to establish a sense of scientific reasoning (Duschl et al., 2007; Ford, 2008).

Vygotsky Social-Constructivist Theory

Vygotsky believed that culture provides children the cognitive tools needed for development (Scott, 2008). Adults such as parents and teachers are channels for the tools of the culture. These tools of culture provides a child the ability to learn through their history, social context, and language (Scott, 2008). The chief theme of Vygotsky's ZPD framework is that social interaction plays a major role in the development of cognition. Vygotsky (1978) stated: "Every function in the child's cultural development appears twice: first, on the social level, and later, on the individual level; first, between people (interpsychological) and then inside the child (intrapsychological)" (p. 57). The interpsychological plane Vygotsky discussed suggests language can be used as a tool for learning (Scott, 2008). Hedegaard (2001) expanded on Vkygotsky's view:

> In Vygotsky's theory, learning is a social process that takes place between people. He conceptualized learning as internalization of social interactions in which communication is central. Learning takes place in social interaction in a specific context which comes internalized by a person. By internalization, Vygotsky did not mean copying but transforming the external interaction to a new form of interaction that guides the child's actions. Internalization does not directly mirror the external social relations; it is a transformed reflection. (pp. 16-17)

It is possible that a teacher's feedback pattern can influence the actions that occur along the interpsychological plane in a classroom. If the teacher's feedback follows a verification pattern then the discourse will primarily be between teacher and student. Asking the student to develop their learning on this social plane may be difficult because

most students do not view their teacher as a peer, but rather a source of knowledge that has already gone through the process of knowledge acquisition (Ford, 2012; Ford, 2008; Scott, 2008). Verification feedback patterns would most likely skip the interpsychological learning stage and ask students to internalize the teacher's dialogue on the intrapsychological stage. Teachers who use verification feedback may ask students to talk in small groups, but if the outcome of the small groups discussion is always trumped by the teachers' authority the motivation to participate in the dialogue may decrease because ultimately their ideas are considered second tier (Hewson, 1992; Hand, Yore, Jagger, & Prain, 2010).

Shute's (2008) elaboration feedback patterns value the interpsychological stage calling for practices that probe, cue, and promote dialogue between students before introducing the content of the discipline. In fact, one of Shute's feedback guidelines that enhances student learning asks a teacher to provide feedback after the learner has attempted to solve the problem with his or her peers (Shute, 2008). This type of feedback would likely build upon the ideas negotiated among social groups (interpsychological) and then ask students to individually reflect upon their experience (intrapsychological).

These two brief examples, elaboration and verification, illustrate how the type of discourse promoted by the teacher may have substantial effects on students as they construct meaning along the interpsychological plane. Leach and Scott (2002) consider that quality science teaching involves students partaking in a public performance on the social plane provided by the classroom. This performance is initiated by the teacher, but

is enacted by students who must write the “script” for the performance and take the lead in “staging” (Leach & Scott, 2002).

For teachers to know how to guide students along these social planes requires knowledge of another aspect of Vygotsky’s social constructivist learning theory; the “Zone of Proximal Development” (ZPD). This component of Vygotsky’s theory, along with how teacher feedbacks influence movement in the ZPD will be discussed in the next section.

Vygotsky's ZPD exemplifies his belief that learning is, fundamentally, a socially mediated activity (Cleborne, Johnson, & Willis, 1997). Vygotsky believed that thinking and problem-solving skills can be placed into three categories: (1) Skills that can be performed independently by the child, (2) skills that can be performed with help from others, and (3) skills that cannot be performed even with help. The actual ZPD would fall in between 2 and 3. The skills are within the grasp of the child’s ZPD if a child uses these cognitive processes with the help of others, such as teachers, parents, and fellow students; they will develop skills that can be independently performed. The types of teacher feedback used as the child moves through the ZPD may have an impact on how easily students make this difficult transition from skills that can be performed with help to skills that can be performed independently. Teachers who use an elaborative approach to feedback may use language as a tool for scaffolding students’ development during this stage.

The term “scaffolding” was first introduced by Jerome Bruner in the 1950s and was a central theme in his book *The Process of Education* (Bruner, 1960). Bruner believed that scaffolding ensured that children aren't left to their own devices to understand something. Similar to the ZPD the support is removed when the student is ready. The scaffolding approach could be viewed differently depending if the teacher uses a deficit or abundance approach in their classroom.

Viewing a student’s skills through a deficit model has been described as looking for areas of weakness in the student’s learning (Harry & Klingner, 2007). The deficit model focuses on skills that the student does not have instead of the opposing abundance model that focuses on skills that the student already possesses. Teachers who view the student through the lens of the abundance model look for skills that the student has and builds upon them to help bridge the gap between those skills and those not mastered yet (Harry & Klingner, 2007).

It is doubtful that teachers who use either verification or elaboration feedback would question the importance of scaffolding during the second stage of navigating through the ZPD. The divergence of their practice would be highlighted by the type of information provided to the student during this stage. A teacher who practices verification feedback may use a deficit model approach and simply “fill in the gaps” for the student once the teacher has acquired data identifying what the student doesn’t know. Conversely, a teacher who uses elaborative feedback practices may use an abundance model approach, evaluating what knowledge the student has acquired and asking probing

questions that catalyze a student into negotiation with peers, and building towards filling these gaps.

In the next section a more detailed description of teacher feedback will be discussed. The categories will fall under the umbrella of *verification* and *elaborative*, but a much more detailed description of each category will be offered. The ZPD will also be re-evaluated in the summary.

Teacher Feedback in Science Classrooms

Attempting to understand and improve students' conceptual understanding of scientific content and the reasoning skills that the students develop has been a major focus in science education over the past several decades. Much of the research has focused on various types of student discourse such as argumentation, explanation, or discussion (Duschl & Grandy, 2008; Duschl et al., 2007; Hardy, Kloetzer, Moeller, & Sodian, 2010; Shute, 2008; Windschitl, Thompson, & Braaten, 2008). However, evaluation of the specific types of teacher feedback patterns in science classrooms has rarely been studied. Most literature in science education over this time has focused on the philosophy of argument-based inquiry (i.e. Jiménez-Aleixandre & Erduran, 2011), peer-peer negotiation (i.e., Ford, 2012), and student outcomes in classrooms that promote scientific dialogue (Keys, Hand, Prain, & Collins, 1999). Little research has evaluated the teacher's actual feedback to students or on what information the teachers are basing their feedback. Teachers' feedback may have an impact on teachers' ability to shape productive classroom discourse by prompting and guiding their students' engagement

(Hardy et al., 2010; Kluger & DeNisi, 1996), and also to foster students' reasoning skills by assisting them in using empirical evidence to support their claims (Hardy et al., 2010; Windschitl et al., 2008). In the next section some specific teacher feedback patterns will be discussed.

Under the larger umbrella of verification and elaboration feedback patterns, another level of discourse is authoritative and dialogic approaches (Shute, 2008). An authoritative approach would fall under the verification pattern and dialogic would follow the patterns of an elaboration approach. Both authoritative and dialogic teacher discourse could be interactive and non-interactive. Definitions and examples of all of the mentioned categories will be discussed next.

Authoritative Discourse

In authoritative discourse the teacher's interventions are intended to transmit information, the emphasis is on the authoritative function of teacher talk (Chin, 2007). A key aspect of authoritative discourse is the teacher's purpose of focusing the students' full attention on just one meaning (Scott et al., 2006). Teachers who use authoritative discourse have little interest in the students' pre-instruction beliefs or conceptual understanding. Instead these teachers focus on beliefs held by the scientific community and their dialogue with the students focuses solely on those beliefs. Bakhtin (1981), described this pattern of discourse as; "The authoritative word demands that we acknowledge it; it binds us, quite independent of any power it might have to persuade us

internally; instead of functioning as a generator of meaning, an authoritative word demands our unconditional allegiance" (p. 342-343).

Even though the nature of authoritative dialogue is to focus students' attention on one view, the discourse between teacher and student can be either interactive or non-interactive. An authoritative/interactive approach might involve teachers focusing on one specific point of view and instructing through a question and answer approach, but with the goal of forming a single point of view (Scott et al., 2006). Teachers using this approach would ask for student opinions and answers, but play "guess what's in my head" games with them, ignoring or rejecting incorrect answers and praising correct responses. The feedback provided to the students still focuses on a single correct answer which is verified and outside ideas are rejected. Teachers who use the interactive/authoritative approach would view their role in Vygotsky's ZPD as a "knowledge filler" using a deficit model to look for what the students don't know (via questioning) and then filling the gaps with direct instruction.

A noninteractive/authoritative approach would be the most extreme pedagogical choice for teachers using a verification framework. In this approach the teacher would simply lecture and not accept questions or comments about the lesson (Hardy et al., 2010). It is difficult to determine how a teacher who uses a noninteractive/authoritative approach would view their role in Vygotsky's ZPD because they would receive no feedback or information about the student's understanding during the lesson. It is possible that the teacher may collect empirical data through standardized assessments and then re-organize future lectures based on the results of the data they collected. In this case

the teacher would also be using a deficit model, looking for the student's missing understanding.

In summary, authoritative feedback, both interactive and non-interactive, gives the teacher the most important voice in the classroom (Chin, 2007). Authoritative dialogue does not ask students to compare and explore each other's ideas (Hardy et al., 2010). In an authoritative classroom ideas or questions which do not contribute to the development of the beliefs of the scientific community are likely to be reshaped or ignored by the teacher. In the next section a contrasting view of teacher discourse, situated within the dialogic framework will be discussed.

Dialogic Feedback

In a classroom that promotes a dialogic discourse the teacher encourages students to put forward their ideas, and to explore and to debate different points of view (Scott et al., 2006). Students in a classroom that uses a dialogic approach might be expected to treat the claims of fellow peers and of themselves as thinking devices (Hardy et al., 2010). Instead of accepting dialogue as information to be passively received and stored, they will instead take an active stance toward it by questioning and extending, and by incorporating the statements of others into their own external and internal utterances (Siddiquee & Ikeda, 2013).

The dialogic framework clearly has connections to Vygotsky's socio-constructivist learning theory. As noted earlier, Vygotsky believed that learning first occurs in a social context on the interpsychological plane. While the authoritative

approach likely skips this plane (or views the interaction of the plane as a teacher-student interaction) and moves directly to the intrapsychological plane, the dialogic approach values the social nuances of learning and uses the ideas and beliefs of students as a starting point upon which to build. Teachers who use a dialogic approach would be more likely to use an abundance model than a deficit model approach. Even though a dialogic approach values student input, it can also be split into interactive and non-interactive categories. These categories will be discussed below.

Teachers who use an interactive/dialogic approach ask students to consider a range of ideas (Scott et al., 2006). Interactive/dialogic teachers pose genuine questions to students as they explore and consider different points of view. Teacher feedback would rarely, if ever, verify whether a student is "right" or "wrong"; instead the teacher would ask students to clarify, generalize, or expand on claims presented to the class (Hardy et al., 2010). Another distinct aspect of interactive/dialogic feedback is teachers welcome and require students to consider different points of view (Duschl & Grandy, 2008). Teachers who use this approach would view Vygotsky's ZPD as a negotiation between students' existing beliefs and beliefs of their peers (Scott, 2008). They would view their role in the process as facilitator who questions and probes students to search for a deeper understanding of everyday phenomena. Teachers who use an interactive/dialogic approach would likely use an abundance model to build upon the student's understanding of the phenomena and help guide it toward a scientific understanding. The larger framework of dialogic feedback seems to require integration, however literature suggests that dialogic feedback can be non-interactive as well.

The key characteristic of dialogic discourse is that it recognizes more than one point of view (Duschl & Grandy, 2008). Teachers who use a noninteractive/dialogical approach would revisit and summarize different points of view, but do so by simply listing them for the student or exploring similarities and differences (Hardy et al., 2010). A teacher may simply start the lesson by listing divergent views of a topic (i.e., some believe global warming is caused by humans, others believe that it is simply nature taking its course) and asking students to research each view. The key difference between interactive and noninteractive dialogic approaches is the voice of the student. In interactive dialogic classrooms the students raise the questions and investigations are based on these questions (Scott, 2008). In a noninteractive dialogic classroom the teacher brings the questions or different views to the students.

Figure 6 provides a summary of the different feedback approaches discussed in the paper thus far. It is likely that an elaborative/dialogic approach is a better fit for Vygotsky's social-constructivist learning theory than is the verification/authoritative approach. The elaborative/dialogic approach allows for student dialogue to be made available on the interpsychological plane. Vygotsky believed that these interactions had to happen first before students attempted to internalize the information (intrapsychological plane). Teacher feedback and student discourse could be used as a scaffolding tool to help students move along the ZPD and reach the third level where skills are recalled with no help. In the next section the focus of the paper will shift towards published research on the effectiveness of teacher feedback on student outcomes.

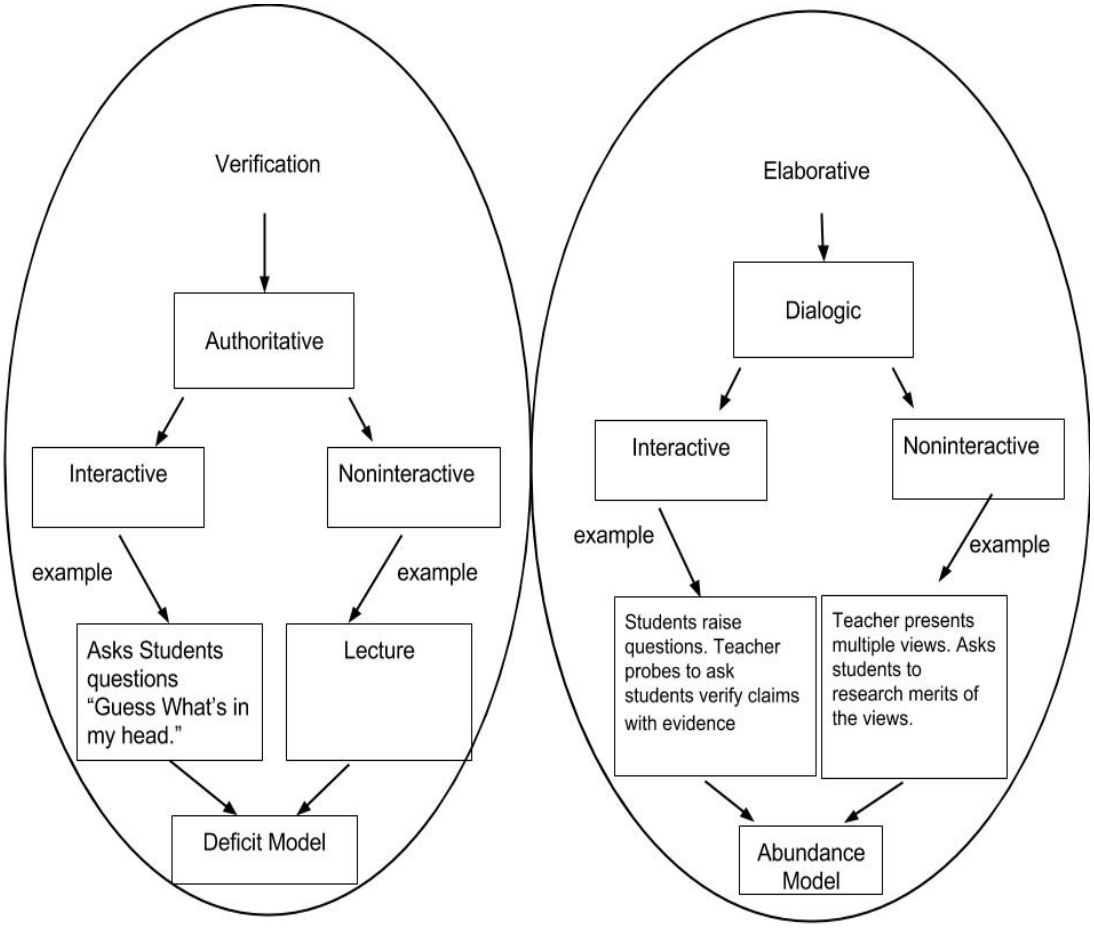


Figure 6 – Examples of teacher feedback patterns and models.

Teacher Feedback and Student Outcomes

Morgan (2006) compared feedback to a good murder because effective feedback depends on three things: "motive (the student needs it), opportunity (the student receives it in time to use it), and means (the student is able and willing to use the feedback)" (p. 76). Before feedbacks are discussed any further a reasonable question to ask is: How influential are feedbacks on student learning? Hattie (2009) offered an answer when he

conducted one of the largest reviews of meta-analyses ever published reviewing over 500 meta-analyses from 180,000 studies, representing approximately 30 million students. Hattie (2009) concluded that feedback had twice the average effect size as typical variables related to student outcomes. Hattie (2009) ranked teacher feedback as one of the highest influences on student achievement.

In a more recent meta-analysis Hattie and Timperley (2007) reviewed teacher feedback and concluded that effective feedback must answer three major questions asked by a teacher and/or by a student: "Where am I going? (What are the goals?), How am I going? (What progress is being made toward the goal?), and Where to next? (What activities need to be undertaken to make better progress?)" (Hattie & Timperly, 2007, p. 86). These questions will be discussed in greater detail in the following sections.

Where am I Going?

Hattie and Timperly's question focuses on a critical aspect of feedback as the information given to students and their teachers about the accomplishment of learning goals related to the task or performance. According to Hattie and Timperly (2007) teachers need to have a clear understanding of the learning goals and must choose appropriate feedback to reach those goals. If this claim is valid effective science teachers would need to understand that different learning tasks require different reasoning skills. Science reasoning has two basic principles: Inductive reasoning (constructing claims with evidence) and deductive reasoning (explaining or predicting a phenomenon with models; NRC, 2012). Lessons at the beginning of the unit may require teachers to use more

inductive prompts to get students to think about their observations, look for patterns, or generate hypotheses. Later in the unit, teachers may move the lessons in a deductive direction by prompting students to think about related models, theories, or explanations that could be generalized.

How am I Going?

Hattie and Timperly (2007) also recognized that teachers' feedback must help students realize where they are as they acquire the appropriate skills to complete tasks. They caution that, "Too often, attention to this question leads to testing, whereas this is not the fundamental conception underlying this question" (Hattie & Timperly, 2007, p. 101). Assessment in the form of tests is one method used by teachers to address this question and it often fails to convey feedback information that helps teachers and their students to know "how they are going" (Hattie & Timperly, 2007). In a science classroom teachers could collect information about how the student is *going* using inductive reasoning prompts that ask whether students' observations fit any patterns or deductive reasoning prompts about how well students' ideas fit with the beliefs of the scientific community.

Where to Next?

Hattie and Timperly's (2007) third question prompts teachers to consider when evaluating their feedback asks where their feedback is directing the lesson. Teachers should not consider simply adding more content when thinking about where to take the direction of the lesson, but instead "The power of feedback, can be used to specifically

address this question by providing information that leads to greater possibilities for learning" (p. 103).

In a science classroom teachers' feedback could possibly support students' ability to connect their current understanding to a scientific understanding by scaffolding through the inquiry process. This scaffolding process could improve the students' understanding of the nature of science and improve their overall scientific literacy.

The three questions, "Where am I going?," "How am I going?," and "Where to Next?," that develop effective feedback provided by Hattie and Timperly (2007) may be more effectively answered by using a pedagogical choice that is dialogic in nature. Hattie and Timperly (2007) claimed that, "feedback has no effect in a vacuum" (p. 82) and the learner must "confirm, add to, overwrite, tune, or restructure feedback in memory" (p. 83).These statements do not align with an authoritative framework that focuses on one possible answer and point of view. A dialogic approach that asks students to consider many possible outcomes would provide the students the opportunity to consider, reject, or modify the feedback provided by the teacher or peer. Teachers may have a blueprint for their units that ends with an authoritative conclusion of what the scientific community currently believes. However, teachers who use a dialogic approach believe that the steps to reach the views of the scientific community involve students constructing their beliefs along the interpsychological plane.

Summary of Feedback

Most feedback studies have focused on students' psychological changes such as motivation, efficacy, or self-regulation (Hattie & Timperley, 2007; Shute, 2008). Research on how teachers' feedback promotes scientific discourse to enhance students' reasoning skills has rarely been studied. From this standpoint of discourse, learning science should not be just accumulating scientific knowledge, but should also include the ability to engage in scientific discourse that fosters students' ability to organize, develop, and evaluate knowledge corresponding to scientific standards (Windschitl et al., 2008).

Helping students develop thinking skills through argumentation has been a widely supported goal in recent science education literature (NRC, 2012, NGSS Lead States, 2013; Wallace et al., 2004). These documents have promoted teaching practices that move beyond experiments and investigations and towards practicing science argumentation. Scientific argumentation, as described in this literature review, can be classified as a process where students engage in a dialogical process where claims are submitted and evidence is provided for those claims. Students also recognize that their claims will be critiqued by their peers and the teacher in an attempt to further the entire classes conceptual understanding of the science taught. A consensus of what is accepted by the scientific community is negotiated by students through listening, reading, writing, and talking (Duschl et al., 2007).

The interactions described above involve a type of teacher/student dialogue that occurs through the use of language. In the scientific community language is the means by

which arguments are constructed, critiqued, and consented to toward understanding its questions, claims, and evidence (Osborne, Erduran, & Simon, 2004; Windschitl, et al., 2008). Using language as a learning tool creates an opportunity where learning science is not only about how to define or label words to explain content or concepts; but rather about the ways in which dialogue can be used to extend students' conceptual understanding of science. Unfortunately, engaging students in science argumentation rarely occurs in classrooms (Driver, Newton, & Osborne, 2000; Lemke, 1990). As teachers attempt to enact an argument-based inquiry approach in their classroom, they are faced with the task of teaching students skills of amassing, presenting, discussing, and critiquing scientific knowledge. Additionally, complications arise because these skills, when used in science argumentation, are not developed in isolation (Jimenez-Aleixandre, Rodriguez, & Duschl, 2000). If students learn to practice effective science argumentation, they gain new skills of how to use language to explain how, and what they know and in doing so, achieve a greater understanding of the nature of science (Norris et al., 2007). Research conducted in the last decade supports the claim that science argumentation should be a core practice in science classrooms (Berland & Reiser, 2009; Jimenez-Aleixandre & Erduran, 2008; Hand, 2009; NGSS Lead States, 2013; NRC, 2012).

CHAPTER 3

METHODOLOGY

This study investigated the relationship between teachers' use of dialogic feedback and the teacher knowledge of their learners on student outcomes on the science section of the Iowa Assessments. The research design used a non-experimental, correlational design (Johnson, 2001) to examine the relationships of two predictor variables, teacher judgment accuracy (JA) and dialogic feedback (DF), and the outcome variable of student achievement on the science section of the Iowa Assessments in grades 3-8 for the 2014-2015 school year (IASci).

Participants

Thirty-three third- through eighth-grade teachers in two moderate-size school districts in the Midwest United States served as participants in the study. The teachers range from two to 34 years of teaching experience. All of the teachers in the study regularly teach science; however, the third and fourth grade teachers teach self-contained classes and are required to teach all academic subjects. The fifth and sixth-grade teachers teach science and math, and the seventh and eighth-grade teachers exclusively teach science. All teachers teach the same students for the entire school year. The number of teachers for each grade level is as follows: eight, third-grade teachers, eleven, fourth-grade teachers, four, fifth-grade teachers, four, sixth-grade teachers, three, seventh-grade teachers and three, eighth-grade teachers.

For anonymity purposes the districts will be referred to as district A and B in this section. The students in district A come from mostly white, middle-class households. Of the 2,203 students 7.7% are considered in a minority group, and 0.3% are labeled English Language Learners. The number of students eligible for free and reduced lunch in District A is =22.6% and a small number of students live below the poverty line (7.7%).

The students in District B are also mostly from white, middle-class households; however District B does have more students in a minority group, are English Language Learners, or receive Free and Reduced Lunch. Of the 1,614 students 15.6% are considered in a minority group, and 2.9% are labeled English Language learners. District B has more students that receive free/reduced school lunch (49.8%) and live below the poverty line (17.1%) than District A does. The variance in student populations was accounted for using an "At-Risk" variable, which will be discussed later in the procedures section.

In accordance with the University of Northern Iowa Institutional Review Board, teachers were invited to participate in the study and were informed that they will be asked to record a 45-minute video of themselves teaching a science lesson. They were also informed that they will be asked to predict how five students in their class would perform on an assessment. Only teachers who read and signed the University of Northern Iowa Institutional Review Board consent form were included in the study. Seven teachers declined to participate in the study.

Materials

Data was collected on two predictor variables: (1) teacher feedback, and (2) teacher judgment accuracy; and an outcome variable of student achievement in science. A description of the materials used and how the data was collected will be used is discussed in the following sections.

Development of a Teacher Feedback Measure

Existing measures. The goal of the teacher feedback data collection was to quantify the teachers' ability to use dialogic discourse patterns with their students in a science classroom. To accurately capture the type of feedback the teachers in the study provide their students, I created a new analytical tool.

When considering alternatives to creating a statistical tool, a search for protocols that measure teachers' ability to conduct reform-based teaching practices was conducted. This search did not yield any protocol that would accurately measure teacher feedback. The published protocols found typically ask researchers to make conclusions about how well the teaching conforms to a pre-identified standard. Examples like the Inside the Classroom: Observation and Analytic Protocol (ICOAP; Weiss, Pasley, Smith, Banilower, & Heck, 2003) include statements that reviewers score on a Likert scale from "not at all" to "to a great extent" and contain statements like: "The teacher had a concrete grasp of the subject matter content inherent in the lesson" (Sawada et al., 2002; Weiss et al., 2003). Issues with subjectivity was why these observation tools were eliminated

from consideration. The considerable amount of subjective decisions the assessor must make during an evaluation (e.g., “Students were reflective about their learning”), and only a few of the items correlate to teacher feedback patterns were issues as well.

A search of other observation protocols found ones that have been developed that label teaching practices without any ruling as to whether the practices are of quality or not. These observation protocols characterize teacher behaviors in the classroom through a series of codes; observers record the frequency of each behavior during a class period (Hora, Oleson, & Ferrare, 2013; West, Paul, Webb, & Potter, 2013). The feedback tool utilized in this study was largely based off the Teaching Dimensions Observation Protocol (TDOP). In this protocol, observers document classroom behaviors in 2-minute sessions throughout the duration of the class period (Hora et al., 2013). The potential classroom behaviors are labeled in 46 codes in six categories, and observers make a tally when any of the behaviors occur.

The TDOP protocol evades the issues of subjectivity associated with the ICOAP, but it only has one section in which observers evaluate teacher-student interaction. This category had only seven components and the codes do not relate well to dialogic feedback patterns. For this reason, a new dialogic-specific feedback tool was created.

Scoring dialogic feedback. To capture the various parts of a science lesson where dialogic interactions could take place the feedback codes were divided into six main categories; those categories are listed below:

1. Development of the views shared by the scientific community;
2. Shifting from an everyday view to a scientific view;
3. Facilitating dialogue;
4. Providing opportunities for reflection;
5. Establishing classroom environment; and
6. Developing scientific understanding.

Under each main heading are three sub-categories creating a total of fifteen characteristics of dialogic feedback (see Appendix A). These main and sub-headings were developed based on the author's expertise in teaching science, coupled with literature found in Chin (2007), Duschl, (2008), Duschl et al. (2007), Ford (2012), Hand (2007), Hardy et al. (2010), Kuhn (2005), Kuhn (2010a), Kuhn (2010b), Kuhn and Crowell (2011), Lemke (1990), Scott (2008), and Scott et al. (2006). The dialogic teacher feedback observation tool has codes developed from science education research on negotiation, scientific argumentation, dialogic feedback, and argument-based inquiry. The tool reduces the likelihood of subjective scoring by identifying teacher behaviors and simply tallying the total behaviors observed rather than speculating on the quality of the behavior.

Like the TDOP, this new protocol documents classroom behaviors in two-minute video intervals throughout the duration of a class session, does not require observers to make judgments of teaching quality, and produces clear quantitative results.

To score the new tool, evaluators watch two minutes of video, and then pause to code behaviors observed during the time span. Total scores are then aggregated at the end of a class session and can be used in statistical analyses.

Validity. Time did not allow for rigorous validation of the feedback observation tool, but two measures of validity were obtained. First, face validity was established by asking two university professors who have expertise in dialogic science instruction to review the instrument. According to Lewis-Beck, Bryman, and Liao (2004); "Face validity is an estimate of the degree to which a measure is clearly and unambiguously tapping the construct it purports to assess" (p. 215). One of the reviewers claimed that, "The tool does a nice job capturing aspects of dialogical teaching that would actually occur in a classroom. I think it matches the literature on dialogical feedback, and presents the codes in a way that can be measured through observation." The second evaluator agreed that the codes used in the feedback tool would likely capture examples of dialogic feedback one would expect to observe from teachers who use an argument-based inquiry approach in science.

Field trial. Data for the second form of validity was collected through a small field trial where six, 45 minute videos were obtained from six elementary teachers. The six teachers were selected using purposeful sampling. The first group of three teachers (two second grade and one first grade) had previously been awarded the Presidential Award for Excellence in Math and Science Teaching (PAEMST; all three were science recipients). The PAEMST is the highest award bestowed to a math or science teacher in

the United States. The author was able to obtain the videos by contacting the teachers and asking permission to view the video they submitted for the PAEMST award.

The second group of teachers (two second grade and one first grade) were teachers at a rural elementary school in the Midwest United States. The teachers have not received any awards or recognition for teaching science and have not attended professional development specializing in inquiry science teaching. Results of the field trial are below in Figure 7.

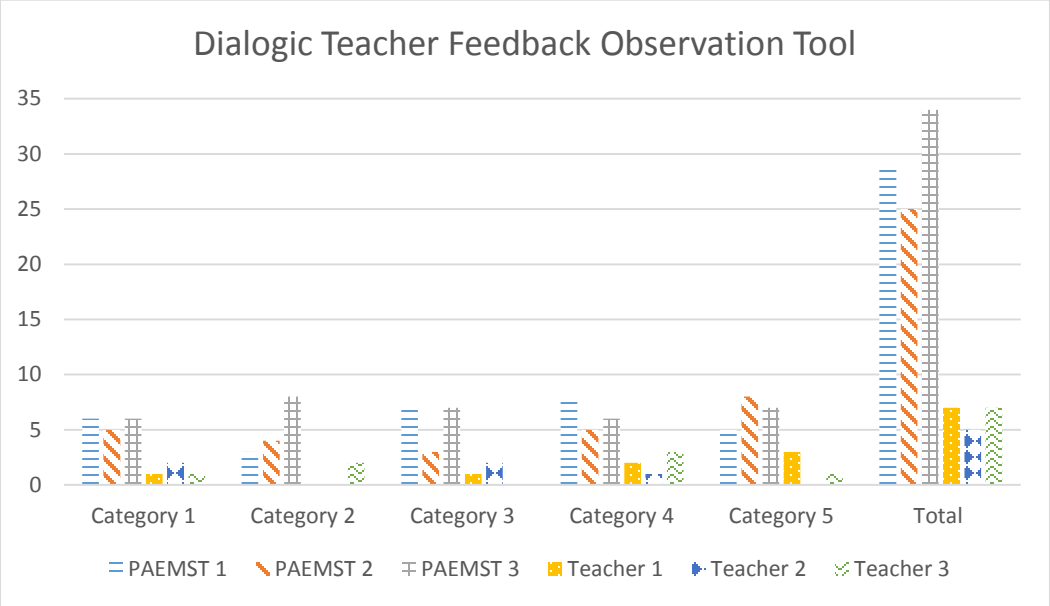


Figure 7- Results from field trial. The three award winning teachers are PAEMST 1-3. The non-award winning teachers are teachers 1-3.

The sample size of this field trial was too small to conduct non-parametric or parametric statistics so a detailed analysis was not performed. However, the teachers who won the PAEMST scored much higher on the feedback observation tool than the teachers

who have had no professional development or experience with inquiry-based science teaching. The videos of the three PAEMST winners all taught different lessons and they recorded their lessons well before the tool was created. The divergent scores between the two groups, and the high marks of the PAEMST cohort suggest that the award winning elementary science teachers use feedback that promotes dialogic teaching practices, providing evidence of the construct validity of this instrument. None of the teachers used in the field trial participated in the study.

Interrater reliability. Reliability of the dialogic feedback tool was established through interrater reliability (IRR). Videos submitted by the participants were scored by myself and a second observer. The second evaluator recruited for IRR has eight years of teaching experience, used a dialogic teaching approach in his science classroom for all eight years, and has attended and worked as a consultant in multiple professional development workshops where the goal was to develop argument-based inquiry teaching among the participants. This individual and I met in the spring of 2015, and I trained him to score the videos. Next, the second evaluator scored one PAEMST video and one non-PAEMST video used in the field trial. The scores for both evaluators on the two videos had a Pearson's $r = .921$, indicating a high level of correlation between the two evaluators' estimate of the teachers' feedback, as measured by the tool used in the study.

Data Collection

Video prompt. Potential subjectivity of the scorer is not the only threat to the internal validity of the feedback measure. A second issue is the type of video that the

teachers submit, and how the prompt given to the teacher may influence the type of lesson taught for this study. To control for this potential threat, a specific prompt was given to the teachers before they recorded their video. To attempt to capture a typical science lesson and not something outside the teachers' regular scope of practice the following prompt was given: "Record a typical science lesson where you are teaching a new concept with the students."

The wording in the prompt was carefully chosen in an attempt to eliminate biased videos from being submitted by the teachers. In the next section the rationale for choosing specific words in the prompt follows. First, the word "typical" was chosen to indicate that the lesson in the video should not capture instructional decision making that is altered or different from the teacher's usual practice.

Second, the words "you are teaching" were selected to indicate a desired video will show the teacher engage in instructional practices where they are involved in the instruction. The word "teaching" was selected over "interacting" because the teacher has the freedom to select a teaching approach of his or her choice and are not forced to select a specific type of teaching that may produce biased results. Asking teachers to submit a video of them "interacting" with the students could instead lead them to submit a video that captures practices that are not typical. For example, teachers may consider a lecture as "non-interacting" and submit a video that shows them attempting a more dialogic interaction with their students. However, if the teacher typically uses a lecture format, then the video and the score they receive on the feedback observation tool would represent an inaccurate view of their typical practice.

Third, the word “new” was selected because the prompt may give the evaluator a glimpse into the teacher’s view of how to introduce new concepts to students. Teachers who use an authoritarian approach may use a deficit model when considering how to effectively teach new concepts to students. They may be inclined to use lecture to “fill” the students’ lack of understanding and then later use investigation as a way to “prove” the lecture. This type of teacher/student interaction is not common in dialogic teaching. Instead teachers using dialogic teaching would likely ask questions, promote discussion amongst the students, build upon their ideas, and investigate questions that they generated (Chin, 2007, Hand, 2009).

Finally, the words “with the students” were selected in an attempt to capture a video of the teacher engaging with their students. As mentioned earlier, prompts like a “good” lesson were eliminated because it may yield videos of student-to-student interaction. By including the word “with” the prompt may cue more interactive behavior without explicitly asking for a specific type of communication.

Data analysis. Each teacher submitted a 45 minute video of a science lesson during a three week window in the spring of 2015. Each video was analyzed using the teacher feedback tool developed by the author. Each teacher recorded the lesson using their district-issued iPad. The videos were then shared with me using a secure website. After an acceptable IRR score was established I scored each video by watching 2 minutes, pausing, then marking whether the teacher’s feedback provided to his or her students matched the categories and items in the feedback tool. Aggregate scores from the feedback measure were then used in the regression analysis. In an attempt to keep the

lesson that the teachers' recorded similar a prompt was given. The rationale of the wording of the prompt will be discussed in the next section.

Interrater reliability. The same second observer mentioned earlier and I watched a sample of videos and compared scores on the evaluation tool. A random sample of fifteen videos (33.3%) were selected for both the author and the volunteer to evaluate and a Pearson's r = .935 was calculated, demonstrating strong IRR between the two evaluators.

The previous section explained in detail how the measure of the predictor variable of teacher feedback was developed and was collected. In the next section a description of how the measure of teacher judgment accuracy was collected is discussed.

Teacher Judgment Accuracy Measure

Teachers were asked to predict their students' outcomes on a nine-question multiple-choice quiz. The author met with a representative from each grade level and discussed what science content was taught in the previous weeks leading up to the quiz date. The timing of the assessment was critical and required coordination between the teachers and the author. It was imperative that the teacher gave the assessment after the lessons in the unit were taught and before any similar assessment had taken place.

After meeting with the teachers to determine what content was taught prior to the JA task, the quizzes were created by the author and a consultant employed as a faculty member at a university and former science assessment developer for an international testing organization. The questions on the quizzes were separated into three groups: (1) three questions on basic content; (2) three questions on phenomenological understanding,

and (3) three questions on rule-based understanding (examples of each can be found in Appendix B –D).

Three levels of questions were selected because each requires a different type of understanding of the content taught. If all nine questions are simple recall questions, the teachers' JA score may not reflect a true picture of their understanding of the students' grasp of the scientific concepts. The three levels of questions were modeled after the assessments developed in Ayala et al., (2008), Herman and Choi (2008), and Hardy et al., (2010). The levels of questions are discussed below.

The first level of questions included basic content level items. These questions required the lowest level of thinking (remembering) on the revised Bloom's Taxonomy (Krathwohl, 2002). Questions at this level did not require a demonstration of a claim backed with evidence, conceptual understanding of the science taught in the unit, nor high-level scientific reasoning. This level of question assessed students' ability to simply recall content that was taught during the unit. An example of a basic content level question was: "What is the formula for force?" (See Appendix B).

The next group of questions evaluated student understanding of science content specific to an observed occurrence in nature. During the pre-assessment meeting, I confirmed that each teacher had taught the science content so the questions were relevant to the students. The questions in this group required slightly more cognitive labor than basic level questions and asked students to explain and interpret the scientific nature of a

single phenomenological occurrence. An example of a phenomenological-level question was: “Why does a heavy rock sink when placed in a pool of water?” (See Appendix C).

The third type of item in the assessment was rule-based questions. These types of questions went beyond explaining science at a phenomenological level and assessed the students’ understanding of the established rule of science for the observations gathered. Rule-based questions determine whether the student can generalize beyond the phenomenological observation and construct patterns that fit an established scientific law. An example of a rule-base question was: “Why do temperatures in January vary in different parts of the country?” (See Appendix D).

All of the students in the classroom took the quiz, but the teacher predicted the outcomes of only five students in his or her class. I generated a pool of student candidates by obtaining student data from the district’s curriculum directors. Students were included in the pool if they had no previous identifiable label that might influence the teacher’s prediction of their academic performance. The students with the following labels were not eligible for selection for the teacher JA task: Individual Education Plan, and Gifted and Talented.

Once the pool of students was created, five students were randomly selected from each class. Next, teachers received copies of the quiz at the end of the day prior to the day they gave the assessment. They predicted how all five students scored on every item on the assessment and the predicted scores were recorded. The next day, the teachers had their students take the assessment and all tests were returned for evaluation.

The JA task in this study was the item-specific hit rate. Teachers were asked to complete the hit-rate prediction task for each of the five selected students for the nine questions on the assessment (See Appendix E). After data were collected for each predictor variable the quantitative analysis was conducted. Details of the analysis are discussed in the next section.

To compute a hit rate score for the JA variable, the number of correct predictions out of the total number of predictions made was calculated. Teachers made predictions about five students' performance on a nine-item quiz, making a total of 45 predictions in all. The teacher received a one point credit when student performance (correct or incorrect) on an item and teacher prediction of that performance matched.

Outcome Variable – Student Achievement

The outcome variable collected in this study was the Normal Curve Equivalent (NCE) score on the science section of the Iowa Assessments for the 2014-2015 school year. NCE scores were obtained by taking the National Percentile Rank (NPR) of each student in the teachers' class and then converting that score to a NCE score. NCEs are similar to percentiles in that they have a mean of 50 and a range from 1-99 (Mertler, 2002). When the rankings of percentiles and NCEs are placed on a normal curve, it is common to observe percentile rank scores to cluster around the mean (50) while the NCE scores are evenly distributed throughout the curve (Mertler, 2002).

After the NCE scores were collected, the mean score for each teacher's students was calculated. This mean score served as the outcome variable data point for the

teachers in the study and the previously discussed predictor variables were used to conduct a regression analysis.

Methods

The respective analyses used to answer the three research questions will be described in this section. All three questions involve regression analyses, however each question required a specific method.

Throughout the analysis $R^2_{adjusted}$ was used instead of multiple R^2. $R^2_{adjusted}$ provides a more conservative estimate of the amount of variance the predictor variables account for in the DV (Miles & Shelvin, 2001). A major difference between R^2 and $R^2_{adjusted}$ is that R^2 assumes that every variable added to the equation explains the variation in the dependent variable, while the adjusted $R^2_{adjusted}$ indicates the percentage of variation explained by the variables that actually affect the dependent variable (Lane, 2008).

Research Question 1

To answer the question about how much variance each predictor variable accounts for independently, a simple hierarchical regression analysis was conducted. Hierarchical regression is a commonly used method to analyze the effect of a predictor variable after controlling for other confounding variables (Pedhazur, 2007). This "control" is acquired by calculating the change in the $R^2_{adjusted}$ at each step of the analysis, thus accounting for the increase in variance after each variable is included in the regression model (Pedhazur, 2007).

Another important aspect of a hierarchical regression is the order of variable entry. Predictor variables are typically added to the equation with regard to their logically determined priority (Miles & Shelvin, 2001). For example, if AtRisk is entered first, and JA second, the change in $R^2_{adjusted}$ will provide a more accurate description of the amount of variance that JA accounts for. In this case, entering AtRisk first allowed the percent of AtRisk students to be accounted for in the JA variable. Two separate analyses were conducted; first JA with AtRisk accounted for, and second DF with AtRisk accounted for. The analyses were conducted independently of each other because the first research question investigates if the two predictor variables predict student outcomes by themselves. Interaction effects between the two variables were investigated in questions 2 and 3.

Research Question 2

To investigate the potential relationship between the predictors, a moderator analysis was conducted to determine if any interaction effects occurred between JA and DF, and if this interaction had any predictive power on IASci.

A moderator analysis is used to determine whether the relationship between two variables depends on (is moderated by) the value of a third variable (Edwards & Lambert, 2007). A model for moderating relationships can be drawn as a linear relationship in which a predictor variable is strongly correlated to another and a second variable (the moderator), alters the strength of that relationship. The model for the moderation analysis is shown in Figure 8.

A moderator analysis was conducted following the standard procedures outlined by Cohen and Cohen (1983) to test the hypothesis regarding the moderating role of JA in the link between DF and IASci. Guidelines provided by Baron and Kenny (1986), and (Frazier, Tix, & Barron, 2004) regarding the use of hierarchical multiple regression analyses to test for moderator effects were followed. Procedures for analyzing and interpreting the interaction terms, recommended by Aiken, West, and Reno (1991), were also employed. All predictor variables were centered following recommendations by Aiken et al., (1991) to reduce multicollinearity between the interaction terms.

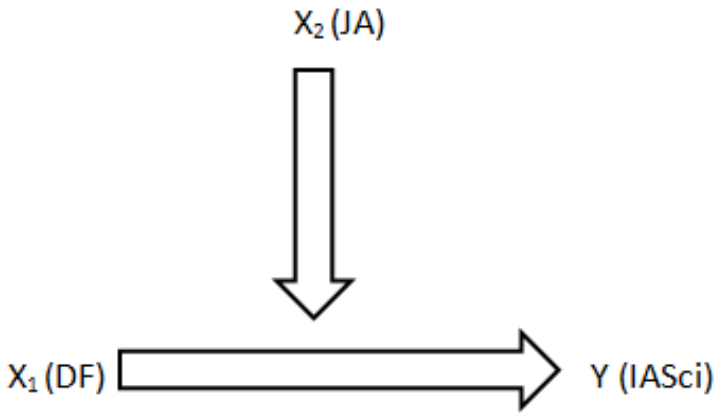


Figure 8 – Model for the moderator analysis based on Baron and Kenny (1986).

The example in Figure 8 shows an arrow from X_2 that points toward the X_1, Y regression path. This path diagram represents the notion that the coefficient for that route is modified by X_2 (Edwards & Lambert, 2007).

Before the analysis was conducted an interaction term was added to the original model used in the first research question. Including an interaction term is a way to statistically account for how a predictor variable has a different effect on the outcome depending on the values of another predictor variable (Edwards & Lambert, 2007). The most common way this interaction is measured in psychological analyses is to calculate the product of the two standardized predictor variables and include them in the multiple regression. The model for this analysis is listed below:

$$Y' = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 * X_2) + \varepsilon$$

The most commonly used approach to conduct a moderator analysis, in psychological research, involves a hierarchical regression that follows these steps: First, run a multiple regression model predicting the outcome variable Y from both the predictor variable (X_2) and the moderator variable (X_2). The ANOVA and summary table should show that both variables and the model in general should be significant. Next, the interaction term, (X_1 * X_2) is added to the original model and a second regression analysis is run. Moderation is occurring if the summary table indicates the model and the interaction term are still significant and there is a significant change in the $R^2_{adjusted}$ score of the two models (Miles & Shevlin, 2001, pp.186-187). Additionally, if the predictor and moderator variables are

not significant once the interaction term is added, then complete moderation has occurred (Miles & Shevlin, 2001).

Research Question 3

The third research question investigated whether mediation effects occurred between the two predictor variables. In regression analyses, predictor variables can have mediation effects on each other. Mediation analysis investigates the mechanisms that underlie an observed relationship between a predictor variable and an outcome variable and examines how they relate to a third variable, the mediator (Hoyle, 1999). The relationships between the predictor, the mediator and the dependent variables can be depicted in form of a path diagram/model (Figure 9). The mediator analysis was chosen because it was hypothesized that the use of dialogic feedback in a classroom may allow teachers access to student knowledge that does not typically present itself in a lecture-type format. In that case the DF may be mediating the student knowledge that the teacher receives due to a student feeling free to express their personal understanding of the science content.

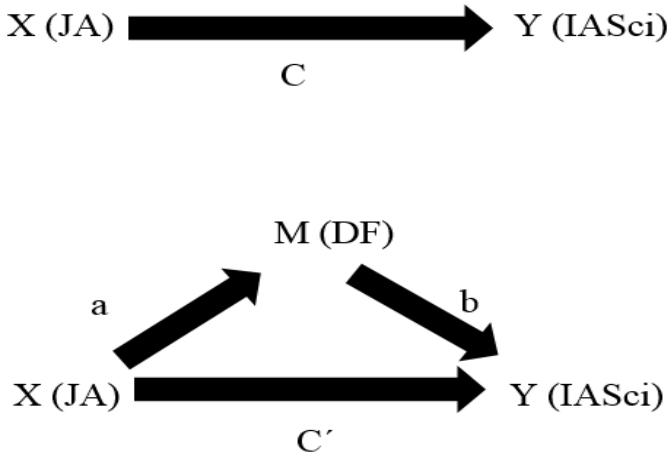


Figure 9 – Model for the mediation analysis. The above path (C), must be established first, then paths a and b determine if path C′ is occurring.

According to Baron and Kenny (1986), four conditions must be met for variable M to be a mediator:

1. X (predictor) is significantly associated with Y.
    a. (Path c) Bivariate regression Y (IASci) = $\beta_0 + \beta_1 X_1$ (JA) + Ɛ
2. X (predictor) is significantly associated with M.
    a. (Path a) Bivariate regression Y (DF)= $\beta_0 + \beta_1 X_1$ (JA) + Ɛ
3. M is significantly associated with Y (after controlling for X).
    a. (Path b) Multiple regression Y (IASci) = $\beta_0 + \beta_1 X_1$ (DF) + $\beta_2 X_2$ (JA) + Ɛ
4. The impact of X on Y is significantly less after controlling for M (full mediation occurs when X = *p* =.00; partial mediation occurs when the *p* value is less).
    a. Path C′

Summary

The purpose of this study was to examine the links among teachers' knowledge of students' understanding of scientific concepts, the degree to which teachers provide dialogic feedback, and students' science achievement outcomes. Teachers' knowledge of science learners was measured using a teacher judgment accuracy task. Teachers in the study predicted how well their students performed on specific items on a science assessment. The type of feedback teachers provide was measured using an observational coding scheme. Teachers were asked to provide a video of a science lesson they have taught, and their feedback to their students in this lesson was coded using an observation tool developed for the study.

CHAPTER 4

RESULTS

This study investigated the relationship between teachers' use of dialogic feedback and the teacher knowledge base of knowledge of learners, reflected in their judgment accuracy, on student outcomes on the science section of the Iowa Assessments.

The study employed a non-experimental, correlational design (Johnson, 2001) to examine the relationships of two predictor variables (teacher judgment accuracy and teacher feedback) and the outcome variable of student achievement on the science section of the Iowa Assessments in grades three through eight for the 2014-2015 school year. In addition to the two predictor variables of interest, an "At-Risk" variable was created in an attempt to account for the academic variance in each classroom. The AtRisk variable was included in the analysis of Research Question 1 because the non-experimental design used in this study did not allow for proper controls used in experimental design. Since teachers could have been assigned a large number of high or low academic performing students when the classroom rosters were created, I thought it was important to attempt to control for the academic variability of each class, by including the AtRisk variable. This variable will be discussed in greater detail in the following section.

"At-Risk" Variable

Time did not allow for a pretest to be given to the students in the study, so an "At-risk" variable was created in an attempt to control for the academic diversity in each classroom. In addition to providing student achievement scores, the curriculum directors

of the two school districts in the study shared coded information indicating whether or not each student in the study had an Individual Education Plan (IEP) or was eligible for the Free/Reduced Lunch Program (FRL). Previous research suggests that students who have either of these labels are likely to perform below their non-labeled peers on standardized tests (e.g., Gersten & Dimino, 2006; Mulhall, Flowers, & Mertens, 2002). The at-risk variable was calculated by using the available data to determine the percentage of students in the teacher's classroom who meet the criteria of an at-risk student (Note: if a student was eligible for FRL and had an IEP they were counted twice).

Tests of Assumptions

Quantitative analyses rely upon certain assumptions about the variables used as predictors. When these assumptions are not met the results may not be accurate, resulting in a Type I or Type II error, or an over- or under-estimation of significance (Keith, 2006). Therefore, the assumptions of linearity, collinearity, and normality were tested before further analyses were run.

All analyses of assumptions and subsequent analyses were conducted using standardized data. Standardizing the variables makes the interactions easier to interpret, because the predictor variables were not measured in the same units. By standardizing the raw scores of the variables, the relative importance in the predictor variables' effect on the dependent variable can be interpreted with more confidence (Keith, 2006).

An examination of correlations (see Table 3) revealed that no predictor variables were highly correlated with each other.

Table 3 – *Correlation matrix and significance codes*

	JA	DF	IASci
DF	.122		
IASci	.358*	.633***	
AtRisk	-.138	-.341	-.148

Note: '***' = $p <$ 0.001; '**' = $p <$ 0.01; '*' = $p <$ 0.05

Further, the predictor variables correlated with the DV at a moderate to high level (see Table 3). The collinearity statistics (i.e., Tolerance, VIF, skew, and kurtosis; see Appendix F) were all within accepted limits. Residual and scatter plots indicated the assumptions of normality, linearity and homoscedasticity were all satisfied (See Figure 10).

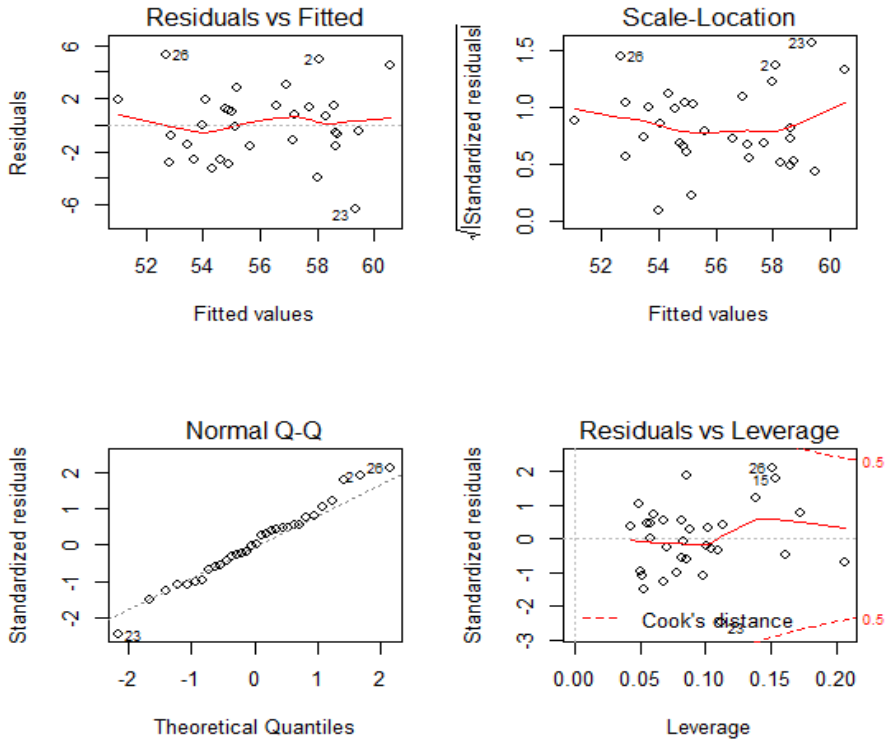


Figure 10 Graphs testing for normality, linearity and homoscedasticity.

According to Figure 10 the residuals vs. fitted graph, (and standardized residuals vs. fitted- the upper right and left graphs) indicate that the residuals and the fitted values are uncorrelated, as they should be in a homoscedastic linear model with normally distributed errors (Osborne & Waters, 2002). The Q-Q plot (lower left) tests the assumption that the errors are normally distributed. In this plot, the points lie close to the dashed line, indicating a normal distribution (Osborne & Waters, 2002). The lower right plot shows the standardized residuals are centered around zero. In normal distribution most points are expected to be symmetrically arranged around zero (Osborne & Waters, 2002). There is a small amount of clustering away from zero, but no major concerns are present.

With analyses indicating that all assumptions for the two predictor variables were met, Research Question 1 was investigated using the following model: Y' = Mean NCE score for the science section of the Iowa Assessments; X_1 = Aggregate score on the AtRisk variable; and X_2 = Aggregate judgment accuracy score (Number of correct predictions on the hit-rate task). For Research Question 1 AtRisk was added to the model first and then JA second. The formula for the final step of the hierarchal regression analysis was:

$$Y' = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

The analysis was then run a second time, using the same model, with the aggregate score on the teacher feedback tool substituted for $X_{2.}$

The model for Research Questions 2 and 3, investigating the interaction effects uses the same model above; however Y' = Mean NCE score for the science section of the Iowa Assessments; X_1 = Aggregate score on the dialogic teaching tool; and X_2 = Aggregate judgment accuracy score. Specifics on how the variables were entered in the model to address moderation and mediation effects will be provided later in this chapter.

In the next section, the study's three research questions will be reintroduced, followed by a short description of the procedures used in the analysis, and the rationale for each question.

Research Questions and Hypotheses

The three research questions that guided this study focus on JA and DF as predictive aspects of quality science teaching. The research questions are as follows:

Research Question 1: What amount of variance does teacher judgment accuracy and dialogic feedback predict, with respect to student outcomes on the science portion of the Iowa Assessments third through eighth grade?

Research Question 2: Does judgment accuracy moderate the relationship between teacher feedback and student achievement?

Research Question 3: Does dialogic teacher feedback mediate the relationship between judgment accuracy and student achievement?

The three research questions have two separate focuses. Research Question 1 asks if the predictor variables, as they were quantified in this study, are good predictors of IASci, as hypothesized in Chapter 2. Research Questions 2 and 3 focus on potential interactions JA and DF may have with each other. The two interactions that were studied using a moderation analysis, looking at any potential moderating effects JA had on DF; and a mediation evaluation, investigating whether DF mediated the relationship between JA and IASci.

Data Analyses

Descriptive Statistics

The students of the 33 teachers in the study had a mean NCE score of 55.64 on the science section of Iowa Assessments (*SD* = 4.12). The scores had a range of a maximum score of 65 and a minimum score of 44 (range = 21). The teachers scored a mean of 33.85 on the JA measure (*SD* = 4.12), with a maximum score of 42 and a minimum score of 27 (range = 15). The teachers scored a mean of 29.21 (SD = 10.54) on the DF measure, with a maximum score of 50 and a minimum score of 8 (range = 42). The maximum percent of AtRisk students in a teacher's classroom was 24 and the minimum was 3% (Range = 21). The mean score for the AtRisk variable was 10.75 (*SD*=10.49).

Research Questions and Results

Research Question 1

The hierarchical regression revealed that at Step 1, AtRisk did not contribute significantly to the regression model (R^2_{adjusted} = -.009, F (1,31) = .689, p =.413), accounting for a negligible amount of the variance in IASci. When JA was added in the second step of the regression model, neither AtRisk nor JA were significant predictors of IASci ($R^2_{adjusted}$ = .080, F (2,30) = 2.403 , p = .108; see Tables 4 and 6). Introducing the second variable (JA) explained an additional 8% of variance in IASci when controlling for AtRisk.

In contrast, when DF was added in Step 2, an additional 36.6% of the variance in IASci and the model was significant ($R^2_{adjusted}$ =.366, F (2, 30) = 10.24, p < .001, see Table 4).

Table 4 – *Change in $R^2_{adjusted}$ in the hierarchical regressions*

Steps	R^2_{adjusted}	$\Delta R^2_{\text{adjusted}}$	Steps	R^2_{adjusted}	$\Delta R^2_{\text{adjusted}}$
DV – IASci IV - AtRisk	-.009	-	DV – IASci IV - AtRisk	-.009	-
DV – IASci IV – AtRisk + JA	.080.	.089	DV-IASci IV – AtRisk + DF	.366***	.375

Note: '***' = p < 0.001; '**' = p < 0.01; '*' = p < 0.05

Table 5–*Step 2 of the DF hierarchical regression.*

IASci ~ zAtRisk + zDF

Coefficients:

	Estimate	Std. Error	*t* value	*Pr*(>|*t*|)
(Intercept)	6.685e-16	1.305e-01	0.000	1.000000
AtRisk	1.098e-01	1.417e-01	0.775	0.444706
DF	6.341e-01	1.414e-01	4.484	0.000106 ***

Note: '***' = $p <$ 0.001; '**' = $p <$ 0.01; '*' = $p <$ 0.05

Residual standard error: 0.7494 on 29 degrees of freedom
Multiple *R-squared*: 0.491, Adjusted *R-squared*: 0.4384
F-statistic: 9.326 on 3 and 29 *DF*, *p* value: 0.0001781

Table 6 - *Step 2 of the JA hierarchical regression*

IASci ~ zAtRisk + zJA

Coefficients:

	Estimate	Std. Error	*t* value	*Pr*(>|*t*|)
(Intercept)	7.309e-16	1.669e-01	0.000	1.0000
AtRisk	-9.984e-02	1.712e-01	-0.583	0.5640
JA	3.443e-01	1.712e-01	2.012	0.0533 .

Note: '***' = $p <$ 0.001; '**' = $p <$ 0.01; '*' = $p <$ 0.05

Residual standard error: 0.9589 on 30 degrees of freedom
Multiple *R-squared*: 0.1381, Adjusted *R-squared*: 0.08061
F-statistic: 2.403 on 2 and 30 *DF*, *p* value: 0.1077

In addition to examining the change in $R^2_{adjusted}$ (see Table 4), Tables 5 and 6 provide information about the predictive values for JA and DF. The data in the regression coefficient table can help us see how a one-unit change in an individual predictor variable, while holding the others constant, predicts outcomes on the DV. This statistical control that regression provides is important because it isolates the role of one variable from all of the others in the model (Miles & Shelvin, 2001). Since JA and DF are continuous variables, the coefficients represent the difference in the predicted value of IASci for each one-unit difference in the variable, if the other remains constant. In this study the variables were measured in points, either on the hit-rate task or score on the dialogic teacher observation tool.

According to Table 5, an increase in one point on the JA measure predicted an increase in the students' NCE scores on the Iowa Assessments by .296, where an increase of one point on the DF measure predicted a .641 increase in students' NCE scores on the Iowa Assessments. These data suggests that the DF variable has a little more than double the predictive power of JA.

In summary, the hierarchical regression indicated that DF was a statistically significant predictor of student outcomes on IASci and JA was not. DF accounts for about five times the amount of variance in IASci than JA. Interaction effects were investigated in Research Question 2 and 3.

Research Question 2

The second research question in this study was based on previous judgment accuracy research that suggests JA, along with other instructional decision making practices, has a positive relationship with student achievement (see, Artlet & Raush, 2014; Behrmann & Souvignier, 2013; Karing et al. (2013). The results addressing the first research question indicate that the two predictor variables (JA and DF) account for roughly 7% and 34.9% of the variance in the student scores on the science section of the Iowa Assessments, respectively. The results of Research Question 1 also indicates that about 3% of the variance is accounted for by a correlation between JA and DF. The AtRisk variable was not included in the moderation analysis, because the focus of the regression was on JA and DF, and the AtRisk variable accounted for almost no variance in the IASci scores.

Research Question 2 investigated potential moderation effects between the two predictor variables. The first step in this approach was already established in the first research question, indicating the model with both predictors included is significant ($R^2_{\text{adjusted}} = .451$, $F\ (2,30) = 14.04$, $p < .001$). Table 7 shows the output of the second step in the hierarchical regression and suggests that moderation has not occurred. When the interaction term was added the model was still significant ($R^2_{\text{adjusted}} = .446$, $F\ (3, 29) = 9.243$, $p < .001$), however the interaction variable was not significant at the < .05 level ($p = .498$). Additionally, the R^2_{adjusted} of the original model (before the interaction term) was .451, and the second model was .446, indicating a $\Delta R^2_{\text{adjusted}}$ of only .005.

Table 7 – *Regression with interaction term added.*

zIASci ~ zJA + zDF + zJA:zDF

Coefficients:

	Estimate	Std. Error	*t* value	*Pr*(>|*t*|)
(Intercept)	0.01343	0.13220	0.102	0.919756
JA	0.27761	0.13421	2.068	0.047616 *
DF	0.57185	0.13908	4.112	0.000295 ***
JA:DF	-0.11326	0.16503	-0.686	0.497953

Note: '***' = $p <$ 0.001; '**' = $p <$ 0.01; '*' = $p <$ 0.05

Residual standard error: 0.7511 on 29 *DF*
Multiple *R-squared*: 0.4800, Adjusted *R-squared*: 0.4459
F-statistic: 9.243 on 3 and 29 *DF*, *p* value: 0.0001894

The lack of a moderating effect suggests that this particular type of interaction between JA and DF did not significantly predict student outcomes on IASci. To further investigate a potential relationship between the predictor variables a mediation analysis was conducted next.

Research Question 3

Results of the steps of the mediator analyses are listed below:

1. X is significantly associated with Y

This condition is satisfied (R^2_{adjusted}= .128, $F(1, 31) = 4.562$, p = <.001)

2. X (predictor) is significantly associated with M

This condition is not satisfied (R^2_{adjusted} = .015, $F(1,31) = .4709$, $p = .498$).

At this point, the mediator analysis was stopped, due to Step 2 not being satisfied. The low correlation between JA and DF nullified any potential change in *p* that would have been found in Steps 3 and 4. A review of the analyses conducted in Chapter 4 are summarized in the section below.

Summary of Quantitative Analyses

A multiple regression analysis was used to test whether two aspects of science teaching significantly predicted student outcomes on the science section of the Iowa Assessments. The results of the regression analysis indicated DF was a significant predictor of IASci ($p < .001$) and the JA variable was not ($p = .053$). An interaction between the two predictor variables was not found as mediator and moderator analyses did not produce statistically significant results. These results are discussed in more detail in the following chapter.

CHAPTER 5

INTERPRETATION OF RESULTS, DISCUSSION, AND IMPLICATIONS

Interpretation of Results

This study investigated the relationships among an aspect of teacher knowledge of their learners (judgment accuracy), teachers' instructional decision making, and student achievement on the science section of the Iowa Assessments. In response to Research Question 1, which addressed the amount of variance teacher judgment accuracy and dialogic feedback predict, the dialogic feedback measure was found to be a statistically significant predictor of the science section of the Iowa Assessments and accounted for a large portion of the variance in the students' scores on that measure. The JA measure was not a statistically significant predictor when the AtRisk variable was statistically controlled.

Statistical analyses related to Research Questions 2 and 3 found that no statistically significant interactions occurred between the predictor variables in the form of moderation or mediation effects. Since there were no effects found in the analyses for Questions 2 and 3, a new theoretical framework was created based on the results reported in Chapter 4 (see Figure 11). The theoretical framework indicates that DF alone is a statistically significant predictor of IASci. JA was not included in the framework because even though it was a theoretically important construct in the original framework it was not a statistically significant predictor of student outcomes, as measured in this study. In

the sections below I will discuss why I think JA was not a significant predictor of student outcomes whereas DF was.

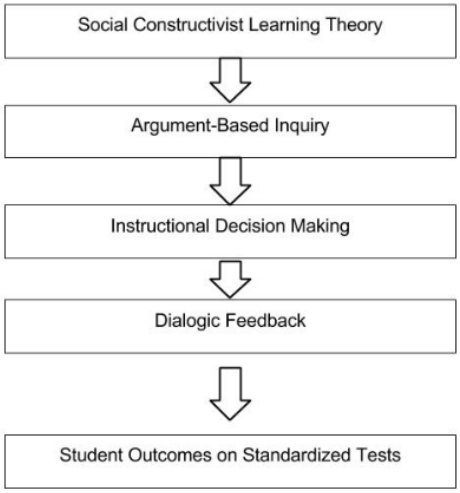


Figure 11- Results; as measured in this study.

Research Question 1

Judgment accuracy. In the Chapter 2 literature review an extensive description of dialogic and authoritative feedback was presented. Figure 6 provided examples of how a teacher may use authoritative and dialogic feedback in the classroom. When the teacher videos were coded for the DF measure, it was clear that the teachers fell into one of the two mentioned feedback patterns. The type of feedback that the teacher gave in the video may provide a clue to how they view knowledge of learners (KOL). For example, Scott (2008) described teachers who use authoritative feedback as "focused principally on an information transmission voice" (p.66). Scott (2008) further described authoritative

teacher utterances as “intended to convey information, and often involved only formal reviews and factual statements” (p. 66). It is possible that the teachers who used authoritative feedback in the videos held the belief of KOL described by Scott (2008). If so, the teachers might view their learners’ knowledge as something that they “gave” to the student, and judgment accuracy would be a measure of how well the students “give” the information back. Assessments, like the one used for the JA measure, would simply be a way to determine how well the students retained the knowledge that was given to them. It is possible that teachers with this epistemological orientation would value standardized assessment because it would be a good measure of how well the students have accepted the content presented to them. In this case, the teacher would not be interested in collecting information about the nature of students’ prior beliefs; instead, he or she would simply want to provide the correct information to the students and expect them to repeat the information back to them on assessments. This view of knowledge, as a static component of learning that can be simply transmitted from teacher to student, is not currently supported by cognitive theorists or science education scholars. Since many of the teachers in the study gave authoritative feedback in their video submissions, and still scored high on the JA task, it may tell us that the standardized nature of the JA task was better suited for these teachers’ ability to predict student achievement, but not for student learning. This may be a reason JA was not a significant predictor of student outcomes when statistically controlled. Teachers who accurately predicted student outcomes on the JA measure were able to demonstrate their knowledge of how students would perform on a standardized assessment. The fact the JA variable was not

statistically significant suggests that being able to predict student responses on assessments that measure convergent thinking, like the JA task did, is not necessarily an aspect of quality science teaching.

Conversely, the dialogic teacher feedback tool measured how teachers promoted dialogue amongst students and helped them shape their thoughts through evaluation of their claims, by providing evidence. It is possible that teachers who provided a lot of dialogic feedback in the videos view KOL as a tool that helps guide instruction. Scott (2008) alluded to this when he described dialogic interactions as having a "generative intent, where outcomes are not pre-determined" (p. 66). Scott (2008) also described dialogic feedback as teacher utterances that were meant to act as "thinking devices where student ideas are generators of meaning" (p.66). This view of KOL may not be easily quantified on a prediction task of a standardized assessment, like the one used for the JA measure. Instead teachers who used a lot of dialogic feedback with their students may be better at predicting more authentic measures of KOL (i.e., student journal writing). Since more than half of the teachers who participated in the study scored above the mean score for the JA measure, and those teachers clearly displayed dialogic feedback patterns consistent with Scott's (2008) description of dialogic feedback, it is possible that these teachers scored lower than what would be expected on the JA measure due to the nature of how it captured KOL. Since the DF measure was intended to quantify how teachers promote dialog amongst their students and asks students to question their personal beliefs and the JA variable measured a prediction of standardized achievement, there could be a mismatch of what the variables are intended to measure. Even though the JA measure

was intended to capture a global picture of the teachers' KOL, the highly specific hit-rate task was based on a sample of five specific students in the classroom. Previous studies that measured interaction effects of JA and an instructional decision making practice used a less specific measure of JA (i.e., Behrmann & Souvignier, 2013; Helmke & Schrader, 1987; Karing, et al. 2013); in those studies, JA was looked at as a general construct of quality teaching and teachers were labeled as either having "high" or "low" JA. In this study, it was hypothesized that a high level of dialogic feedback would provide teachers access to student knowledge that teachers who provide authoritative feedback do not have privy to. However, it is possible that teachers who provided authoritative feedback to their students may have been better predictors on the hit-rate task simply because they have assessed their student in this manner more frequently. It is likely that a standardized assessment, like the one used in the JA task, would be used by teachers who gave authoritative feedback because it is a simply and efficient way to measure student knowledge.

Teachers who value dialogic interactions would likely be more interested in assessing how student ideas change over time. This type of change would be very difficult to capture on a standardized assessment, and would likely be measured through writing or other types of assessment. This mismatch of what the JA task measured and what the DF task measured may be a reason JA was not a statistically significant predictor of student achievement and might explain why there was no interaction between the two variables. In the next section the other component of Research Question 1, dialogic feedback, will be discussed.

Dialogic feedback. The results presented in Chapter 4 show that DF was also a significant predictor of student achievement on the science section of the Iowa Assessments. Much of the science education research, with respect to teachers' instructional approaches, is based on argument-based inquiry that promotes negotiation taking place on a social plane in the classroom. For this reason, the DF variable was chosen as a way to measure teacher instructional decision making, as situated within the argument-based reasoning approach to teaching. In the following sections, an explanation is provided for why DF was a significant predictor of student achievement, referring to the theoretical framework established in Chapter 1.

Over the last few decades research in science education has focused on studies of how meaning is developed through language and negotiation in the classroom. Many of the findings have claimed that scientific knowledge is socially constructed through negotiation (Osborne, et al., 2004). A key element of this negotiation takes place through oral discourse. Group negotiation, including student-to-student and teacher-to-student, is therefore central to understanding how knowledge is created in a science classroom.

The focus of science education research over the last few decades moved much of the science education community away from studies that focused on individual student learning, in isolation, and toward studies that focused on how knowledge is constructed, by students, in a social context in the classroom. This new direction also signaled a rejuvenated interest in the role played by the teacher in the science classroom (Lembke, 1990). Many science education researchers view the teacher's role as a "director" of classroom discourse, encouraging students to challenge fellow classmates and back

claims with evidence; rather than the sole provider of the discourse, where the teacher simply transmits information to the students. The DF tool was designed to capture some of these interactions where the teacher uses language to stimulate and extend students' thinking and advance their learning and understanding. One particular aspect of these interactions that the DF measure captures is language that promotes the use of scientific argument among students.

Previous research has found that engaging in scientific argument enhances students' understanding of scientific concepts, improves their understanding of the science process, and encourages the development of critical thinking skills by making student thinking processes more available (Berland & Reiser, 2009; Ford, 2006; Hand et al., 2010). Zee and Minstrell (1997) described these dialogic interactions as a place where students can express their own thoughts, explanations, and questions, followed by the teacher and students engaging in a prolonged series of questioning exchanges that help students better eloquent their beliefs and conceptions. These interactions include times when students exchange ideas with each other and also when they try to understand the thinking of their peers as they construct an argument (Zee & Minstrell, 1997). The delicate interactions mentioned above require teachers to know how to direct the dialogue of the classroom negotiation; including when to allow students to struggle with concepts and when to direct them to resources that will help them understand the content. The dialogic teacher feedback tool attempted to capture some of these interactions and quantify how often teachers use them during an average science lesson.

Results presented in Chapter 4 indicate that DF is a significant predictor of student achievement in science on the science section of the Iowa Assessments. The use of dialogic interactions forced students to make their understanding of the science content public, and created a situation where students realized that their claims were either backed with solid evidence, they needed to add more knowledge, or they needed to completely change their original ideas. Either way, the use of dialogic feedback as an instructional strategy, as measured by DF, that made students confront their ideas about the science content and self-reflect on the validity of their reasoning. It is hypothesized that asking students to engage in this type of thinking is one of the reasons why DF was a significant predictor of student achievement.

The theoretical model this study was based on suggests that an interaction effect would occur between the two predictor variables. Research Questions 2 and 3 investigated whether a moderator or mediator effect occurred between the predictor variables, and neither was found in the analyses. In the next sections, I will discuss reasons why moderation and mediation effects may not have occurred.

Research Questions 2 and 3

Since Research Questions 2 and 3 both investigated interactions between the variables, yet none was found, it is likely that the lack of statistical significance for both research questions was due to the same issue, so possible explanations for these findings are combined in this section.

The second research question investigated whether a moderation effect was present between the two predictor variables. It was hypothesized that JA was moderating the effect between DF and student outcomes. The assumption of moderation was based on the prediction that if a teacher utilized dialogic feedback, as measured by DF, student outcomes could be predicted accordingly, but if groups of teachers had similar levels of DF, the ones with higher knowledge of their students, as measured by JA, would further predict student outcomes due to the interaction between knowledge of learners and dialogic feedback.

The third research question hypothesized that dialogic feedback mediated the relationship between knowledge of learners and student achievement. The prediction was based on the idea that if teachers used dialogic feedback they would gain access to students' personal interpretations of the science content. Access to this knowledge would allow teachers to make more accurate prediction of their students' performance on standardized assessments, due to the personalized understanding of the science content that the teacher has gained access to.

A potential reason no interaction was found in either analysis may have been due to the type of knowledge the measures collected. The DF measure was created to capture feedback that teachers use to promote negotiation of meaning amongst students that ultimately leads to students constructing scientific claims based on evidence. Steps taken along this journey require repeated stops where teachers personalize instruction and dialogue to fit the needs of their learners. In a way, the DF tool is not only measuring the teachers' use of feedback practices that promote dialogic interactions, it is also measuring

how many opportunities teachers gave themselves to collect authentic knowledge from their students. Teachers who utilized probing questions, asked students to address a peer's comment, or asked a student to clarify a statement gave themselves more opportunities to allow the social exchange of knowledge to occur. These exchanges require students to share personal understandings of content and scientific procedures.

The JA task asked teachers to make predictions on a standardized assessment that does not account for personalized interpretation, but instead asks the teacher for a judgment of convergent thinking. The fact that DF correlated very high with IASci indicates that the use of instructional decisions that promote dialogic interactions predicts student achievement on standardized tests. However, it is possible that the teachers who use dialogic feedback may not have picked up on student knowledge that showed up on the standardized assessment used for the JA task. It is possible that teachers who value dialogic interactions would also be better predictors of a more personalized assessment (i.e., writing essays or other write-to-learn activities).

Most teachers in the study scored fairly high score on the JA task. In the case of a teacher who scored very high on the DF measure it is possible that they obtained enough information to predict scores on the JA task. However, this teacher may have more knowledge of the learner that did not appear on the standardized tests, due to the personalized way that knowledge of learners was acquired via dialogic interactions.

Limitations of the Study

There are several limitations of this study that should be considered. One potential issue with the study was that the teachers who participated came from nine different schools, two different districts, and six different grade levels. However, it is unlikely that this diversity affected the results because the study evaluated aspects of quality science teaching and these characteristics are not reserved for specific grade spans or school districts.

There was additional variance in the amount of time teachers had with students. Elementary teachers in the study were in self-contained classrooms and had the same students all-day, for the entire school year. The fifth through eighth grade teachers have the same students all school year, but only have them half of the day. It was decided that including the fifth- through eighth- grade teachers was acceptable, because without them the study would have serious issues with power. In addition, even though the fifth through eighth grade teachers only have the students for part of the school day, they teach science the same amount (an hour per day), or more, than the elementary school teachers typically teach science.

Another potential threat to the validity of the data was the small sample size used in the study. The relatively small number of teachers who participated in this study means that the results should be generalized with caution. However, Cohen (1988) considered a sample size of thirty-two of sufficient power to detect any effects that may have been present (power =.80, alpha = .05) and this study had thirty-three participants.

The lack of diversity in the students who participated was another limitation of the study. The majority of the students who participated where white and from middle-class families. District B had more minority and FRL students, but these numbers were still very low compared to the rest of the district. Future research is needed to replicate this study in a more diverse setting than the one used in this study. If the study were replicated under ideal conditions the number of participating teachers should be at least doubled. In addition, the student population would be more diverse and before the students completed the hit-rate task a pre-test would be given so the pool of students randomly chosen would be academically similar, based on the results of the pre-test.

Future research on this topic should ask teachers to complete JA tasks of varied specificity, including less specific JA measures like grade equivalency measures that take into account the scores of the entire class and which could be compared to the whole class average of the IASci scores.

Despite the limitations mentioned above the study did produce results that may be useful to various educational stakeholders. Recommendations to these groups will be discussed in the next sections.

Recommendations

The main findings of the study were that DF was a significant predictor of IASci whereas JA was not. However, the JA variable may be of interest to scholars since this study was the first to measure JA as an aspect of quality science teaching. Since this study was the first of its kind to measure JA and DF, in the manner outlined in the

methods section, the data collected may be useful to various groups in the field of education. Since no interaction effects were found, the primary focus of the following sections will be on how researchers and teachers might benefit from the results of Research Question 1.

Recommendations for Scholars

Dialogic feedback. The large amount of variance that DF accounted for in IASci is information that could be useful for science education scholars. As mentioned in Chapter 3, numerous teacher observation tools exist, but many of them require value judgments and very few quantify specific teacher behaviors. Further research is required to investigate the validity of the dialogic feedback tool, but the data collected to answer Research Question 1 suggests that it may be a reliable way to measure feedback that promotes dialogic interactions among students.

Many science education researchers place value on instructional strategies that ask students to articulate their understandings in personal ways as a means to construct a richer conceptual framework of science knowledge. These interactions highlight the collaborative nature of scientific argumentation, where learners are asked to participate in an ongoing series of negotiating and illuminating meanings and descriptions of their knoweldge with their peers and teacher. The teacher feedback patterns in the dialogic teacher feedback tool were designed to quantify these classroom interactions that promote discussion where students' personal explanations and observations are tested against the perceptions and contributions of the broader group and scientific community.

The amount of variance that the DF variable accounted for in IASci was at a very significant level, suggesting that the dialogic teacher feedback tool may be useful to researchers who want to quantify dialogic interactions. Having a way to quantify a teacher's ability to enact dialogic interactions may be useful to researchers as they continue to investigate the effects dialogic feedback has on student learning.

Judgment accuracy. Whereas the findings of this study did not support the predictive value of judgment accuracy in providing dialogic feedback or student science achievement outcomes, JA should still be of interest to science education researchers due to the focus of student-centered curricula in contemporary science research. One way this information could be valuable to researchers in higher education is by emphasizing the topic of teacher knowledge of learners to pre-service teachers. Some researchers investigating pedagogical content knowledge have found that novice teachers tend to rely on content knowledge as a focus of their lessons (e.g., McConnell, Parker, & Eberhardt, 2013; Park & Chen, 2012). If knowledge of student knowledge is an area that receives more attention in science education and educational psychology courses, pre-service teachers may enter the field with a greater appreciation for this knowledge base.

In addition to helping pre-service teachers, the results of the study may prompt further research of JA in science education. A large number of science education researchers use Vygotsky's social constructivism theory as the theoretical base for their research designs. Central to Vygotsky's theory is the Zone of Proximal Development where students negotiate meaning through the inter- and intra-psychological planes. One of the main reasons the JA variable was evaluated in this study was to get a sense of how

accurate teachers were at capturing their students' thought process that was occurring during these interactions. Being able to quantify an estimate of a teachers' ability to gauge their students' thinking as they embark on this complicated endeavor is something that has not been researched in science education and warrants further investigation.

Recommendations for Teachers

Dialogic feedback.. The DF tool could be useful to teachers as they learn new ways to teach science that focus on dialogic interactions between the teacher and students. A common way in-service teachers learn new teaching approaches is through professional development. The DF measure could be used as an assessment tool in professional development workshops attempting to improve teachers' dialogic interactions with students. One of the most difficult aspects of professional development is taking the information learned and implementing it in the actual classroom (Cantrell & Hughes, 2008; Franke, Carpenter, Levi, & Fennema, 2001; Loucks-Horsley, 1998). Contemporary science standards in the United States (i.e., NGSS) place a premium on student-centered negotiation and ask teachers to enact a pedagogy that promotes dialogic interactions. The increased focus on providing students opportunities to develop negotiation skills in science has led to a greater need for quality professional development for teachers. Many current teachers did not participate as students in science learning situations characterized by pedagogy consistent with the methods they are being encouraged to utilize. Consequently, these teachers find calls to change their teaching approach disconcerting because of an uncomfortable feeling with the teaching practices due to their own experiences learning science as a student.

The DF tool could be used by teachers as a self-evaluation tool as they implement a teaching approach that may be, initially, difficult to master. The DF tool captures actual teacher interactions instead of vague, value judgments like other observation protocols. The specificity of these interactions could potentially serve as "training steps" as teachers learn how to blend their personal epistemological beliefs with the teaching approach outlined in this chapter.

Future Research

There are a number of potential follow up studies that could be conducted based on this thesis. First, further studies about potential moderator or mediator effects occurring between the two predictor variables should be conducted. If the study was re-created with a larger sample size and better screening for the JA task (pre-test) the results could be generalized with more confidence.

Additional follow up studies that use a less specific JA measure to determine if any interaction effects occur may also be conducted. It is possible that a more global prediction task (e.g., percentage correct) has the capability to measure a teachers' ability to accurately group students by level of understanding of the science content. The specific attention that the teacher provides each group may be a key aspect of instruction that helps the student learn.

There are also many questions about how teachers gain knowledge of their learners. It would be interesting to interview some of the teachers who had a combination of high DF and JA, and had students who scored high on IASci. Qualitative research may

help find patterns of knowledge acquisition. These patterns could be used to design larger quantitative studies. It would also be interesting to replicate the study and ask teachers to make predictions of assessment that were not as standardized as the one used in this study to see if the scores on the judgment accuracy measure would vary due to the potential access to personalized knowledge of students potentially gained by teachers during the dialogic interactions.

Finally, further research should be conducted on the dialogic teacher feedback tool used in the study. The DF measure was a significant predictor of IASci, and further tests of validity should be conducted to determine if the tool is as effective predicting outcomes in other settings as it was in the study. A high quality observation tool would be a valuable contribution to the science education research community and it would be useful to practitioners who are learning how to include dialogic feedback as a part of their instruction.

As an afterword I wanted to mention, a well-designed dissertation should set a scholar's research agenda as he or she enters the academic community (Bolker, 1998). This study has not only been beneficial as a learning tool on how to conduct quality research, it also has left many questions unanswered and provides an opportunity to contribute to the field of science education and judgment accuracy. The process has been challenging, but I now feel prepared to enter the academic community with the tools necessary to become a successful researcher.

REFERENCES

Aiken, L. S., West, S. G., & Reno, R. R. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage Publications.

Alexander, R. J. (2006). *Towards dialogic teaching: Rethinking classroom talk*. Thirsk: Dialogos. Hong Kong Institute of Education. Hong Kong, China.

Anders, Y., Brunner, M., & Krauss, K. (2011). Diagnostic skills of mathematics teachers and the performance of their students. *Psychology in Education, 3*(2), 175-193.

Anderson, R. (2002). Reforming science teaching: What research says about inquiry. *Journal of Science Teacher Education, 13*(1), 1-13.

Artlet, C., & Raush, T. (2014). Accuracy of teacher judgments. When and for what reasons? In Krolak-Schwerdt, S., Glock, S., & Böhmer, M. (Eds.), *Teachers' professional development: Assessment, training, and learning* (pp. 27-44). Rotterdam, The Netherlands: Sense.

Ayala, C. C., Shavelson, R. J., Ruiz-Primo, M. A., 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.

Bakhtin, M. M. (1981). *The dialogic imagination: Four essays*. Austin: University of Texas Press.

Bangert-Drowns, R. L., Kulik, C. C., Kulik, J. A., & Morgan, M. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research, 61*(2), 213-238.

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*(6), 1173-1182.

Behrmann, L., & Souvignier, E. (2013). The Relation between teachers' diagnostic sensitivity, their instructional activities, and their students' achievement gains in reading. *Zeitschrift Für Pädagogische Psychologie, 27*(4), 283-293.

Benchmarks for science literacy. (1993). New York, NY: Oxford University Press.

Berland, L. K., & Reiser, B. J. (2009). Making sense of argumentation and explanation. *Science Education, 93*(1), 26-55.

Black, P., & Wiliam, D. (2010). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappa, 92*(1), 81-90.

Bolker, J. (1998). *Writing your dissertation in fifteen minutes a day: A guide to starting, revising, and finishing your doctoral thesis*. New York, NY: H. Holt.

Bricker, L., & Bell, P. (2008). Conceptualizations of argumentation from science studies and the learning sciences and their implications for the practices of science education. *Science Education, 92,* 473-498.

Brown, K., & Wilson, D. (2006). *Using Progress Variables And Embedded Assessment To Improve Teaching And Learning.* Paper Presented at the Annual Meeting of the American Education Research Association (AERA), San Francisco, California, April 2006.

Bruhwiler, C., & Blatchford, P. (2011). Effects of class size and adaptive teaching competency on classroom processes and academic outcome. *Learning and Instruction, 21*(1), 95-108.

Bruner, J. S. (1960). *The process of education*. Cambridge, MA: Harvard University Press.

Bruner, J. (1961). The act of discovery. *Harvard Educational Review, 31*(1), 21-32.

Bybee, R.W., Buchwald, E., Crissman, S., Heil, D. R., Kuerbis, P. J., Matsumoto, C. & McInerey, J. D. (1989). *Science and technology education for the elementary years: Frameworks for curriculum and instruction.* Washington, D.C.: The National Center for Improving Instruction.

Bybee, R. W., & McInerney, J. D. (1995). *Redesigning the science curriculum: A report on the implications of standards and benchmarks for science education*. Colorado Springs, CO: BSCS.

Cantrell, S. C., & Hughes, H. (2008). Teacher efficacy and content literacy implementation: An exploration of the effects of extended professional development with coaching. *Journal of Literacy Research HJLR J. of Literacy Res., 40*(1), 95-127.

Chin, C. (2007). Classroom Interaction in Science: Teacher questioning and feedback to students' responses. *International Journal of Science Education, 28*(11), 1315-1346.

Cleborne, M. D., Johnson, D. L., & Willis, J. (1997). *Educational computing: Learning with tomorrow's technologies*. Boston, MA: Allyn & Bacon.

Clermont, C. P., Krajcik, J. S., & Borko, H. (1993). The influence of an intensive in-service workshop on pedagogical content knowledge growth among novice chemical demonstrators. *J. Res. Sci. Teach. Journal of Research in Science Teaching, 30*(1), 21-43.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: L. Erlbaum Associates.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences*. New York, NY: Wiley.

Coladarci, T. (1986). Accuracy of teacher judgments of student responses to standardized test items. *Journal of Educational Psychology, 78*(2), 141-146.

Deboer, G. E. (2000). Scientific literacy: Another look at its historical and contemporary meanings and its relationship to science education reform. *Journal of Research in Science Teaching, 37*(6), 582-601.

Demaray, M., & Elliott, S. (1998). Teachers' judgments of students' academic functioning: A comparison of actual and predicted rating scales. *School Psychology Quarterly, 13*, 8-24.

Dohert, J., & Conoll, M. (1985). How accurately can primary school teachers predict the scores of their pupils in standardized tests of attainment? A study of some non-cognitive factors that influence specific judgment. *Educational Studies. 11*, 41-60.

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

Duschl, R. A., & Grandy, R. E. (2008). *Teaching scientific inquiry: Recommendations for research and implementation*. Rotterdam, The Netherlands: Sense.

Duschl, R. A., Schweingruber, H. A., & Shouse, A. W. (2007). *Taking science to school: Learning and teaching science in grades K-8*. Washington, D.C.: National Academies Press.

Eckert, T. L., Dunn, E. K., Codding, R. S., Begeny, J. C., & Kleinmann, A. E. (2006). Assessment of mathematics and reading performance: An examination of the correspondence between direct assessment of student performance and teacher report. *Psychology in the Schools, 43*(3), 247-265.

Edwards, J. R., & Lambert, L. S. (2007). Methods for integrating moderation and mediation: A general analytical framework using moderated path analysis. *Psychological Methods, 12*(1), 1-22.

Feinberg, A. B., & Shapiro, E. S. (2003). Accuracy of teacher judgments in predicting oral reading fluency. *School Psychology Quarterly, 18*(1), 52-65.

Ford, M. (2008). Disciplinary authority and accountability in scientific practice and learning. Science Education Sci. Ed., 92(3), 404-423

Ford, M. J. (2010). Critique in academic disciplines and active learning of academic content. *Cambridge Journal of Education, 40*(3), 265-280.

Ford, M. J. (2012). A dialogic account of sense-making in scientific argumentation and reasoning. *Cognition and Instruction, 30*(3), 207-245.

Franke, M. L., Carpenter, T. P., Levi, L., & Fennema, E. (2001). Capturing teachers' generative change: A follow-up study of professional development in mathematics. *American Educational Research Journal, 38*(3), 653-689.

Frazier, P. A., Tix, A. P., & Barron, K. E. (2004). Testing moderator and mediator effects in counseling psychology research. *Journal of Counseling Psychology, 51*(2), 157-157.

Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review, 102*(4), 652-670.

Funder, D. C. (2012). Accurate personality judgment. *Current Directions in Psychological Science, 21*(3), 177-182.

Gersten, R., & Dimino, J. A. (2006). RTI (Response to Intervention): Rethinking special education for students with reading difficulties (yet again). *Reading Research Quarterly, 41*(1), 99-108.

Green, J., & Luke, A. (2006). *Review of research in education*. Washington, D.C.: American Educational Research Association.

Gross, A. G. (1990). *The rhetoric of science*. Cambridge, MA: Harvard University Press.

Hackling, M., Smith, P., & Murica, K. (2010). Talking science : Developing a discourse of inquiry. *Teaching Science, 56*(1). 17-22.

Hand, B. (2007). *Science inquiry, argument and language: A case for the science writing heuristic*. Rotterdam, The Netherlands: Sense.

Hand, B. (2009). *Negotiating science: The critical role of argument in student inquiry, grades 5-10*. Portsmouth, NH: Heinemann.

Hand, B., Yore, L. D., Jagger, S., & Prain, V. (2010). Connecting research in science literacy and classroom practice: A review of science teaching journals in Australia, the UK and the United States, 1998–2008. *Studies in Science Education, 46*(1), 45-68.

Hanuscin, D. L. (2013). Critical incidents in the development of pedagogical content knowledge for teaching the nature of science: A prospective elementary teacher's journey. *Journal of Science Teacher Education, 24*(6), 933-956.

Hardy, I., Kloetzer, B., Moeller, K., & Sodian, B. (2010). The analysis of classroom discourse: Elementary school science curricula advancing reasoning with evidence. *Educational Assessment, 15*(3-4), 197-221.

Hargreaves, E. (2013). Inquiring into children's experiences of teacher feedback: Reconceptualising Assessment for Learning. *Oxford Review of Education, 39*(2), 229-246.

Harry, B., & Klingner, J. (2007). Discarding the deficit model. *Educational Leadership, 64*(5), 16-21.

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. London, England: Routledge.

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

Hedegaard, M. (2001). *Learning and child development: A cultural-historical study*. Aarhus, Denmark: Aarhus University Press.

Helmke, A., & Schrade, F. (1987). Interactional effects of instructional quality and teacher judgment accuracy on achievement. *Teaching and Teacher Education, 3*, 91-98.

Herman, J., & Choi, K., (2008), *Formative assessment and the improvement of middle school science learning: The role of teacher accuracy*, Los Angeles: University of California Publications.

Hewson, P. W. (1992). *Conceptual change in science teaching and teacher education.* Paper presented at a meeting on "Research and Curriculum Development in Science Teaching," under the auspices of the National Center for Educational Research, Documentation, and Assessment, Ministry for Education and Science, Madrid, Spain.

Hewson, P. W., & Hewson, M. G. (1984). The role of conceptual conflict in conceptual change and the design of science instruction. *Instructional Science, 13*(1), 1-13.

Hill, H., Ball, D. L., & Schilling, S. (2008). Unpacking "pedagogical content knowledge": Conceptualizing and measuring teachers' topic-specific knowledge of students. *Journal for Research in Mathematics Education, 39*(4), 372-400.

Hoge, R., & Butcher, R. (1984). Analysis of teacher judgments of pupil achievement level. *Journal of Educational Psychology, 76*, 777-781.

Hoge, R. D., & Coladarci, T. (1989). Teacher-based judgments of academic achievement: A review of literature. *Review of Educational Research, 59*(3), 297-313.

Hopkins, K. D., George, C. A., & Williams, D. D. (1985). The concurrent validity of standardized achievement tests by content area using teachers' ratings as criteria. *Journal of Educational Measurement, 22*(3), 177-182.

Hora, M.T., Oleson, A., & Ferrare, J.J. (2013). *Teaching Dimensions Observation Protocol (TDOP) User's Manual*.Madison: Wisconsin Center for Education Research, University of Wisconsin–Madison; 2013.http://tdop.wceruw.org/Document/TDOP-Users-Guide.pdf

Hoyle, R. H. (1999). *Statistical strategies for small sample research*. Thousand Oaks, CA: Sage Publications.

Jimenez-Aleixandre, A., & Erduran, A. (2011). Argumentation in science education: perspectives from classroom-based research. *Science & Education, 20*(5-6), 585-588.

Jiminez-Aleixandre, M. P., Rodriguez, A. B., & Duschl, R. A. (2000). Doing the lesson? or doing science?: Argument in high school genetics. *Science Education, 84*(6), 757-792.

Johnson, B. (2001). Toward a new classification of nonexperiemental quantitative research. *Educational Researcher, 30*(2), 3-13.

Jonassen, D. (1999). Towards a constructivist design model. *Educational Technology, 34*(4), 34-37.

Kahle, J. B., Meece, J., & Scantlebury, K. (2000). Urban African-American middle school science students: Does standards-based teaching make a difference? *Journal of Research in Science Teaching, 37*(9), 1019-1041.

Karing, C., Pfost, M., & Artelt, C. (2013). Is secondary school teacher judgment accuracy related to the development of students' reading literacy? In Pfost, M.,Weinert, S., & Artelt, C. (Eds.), *The development of reading literacy from early childhood to adolescence* (pp. 279-311). Bamberg, Germany: University of Bamberg Press.

Keith, T. (2006). *Multiple regression and beyond*. Boston, MA: Pearson Education.

Keys, C. W., Hand, B., Prain, V., & Collins, S. (1999). Using the science writing heuristic as a tool for learning from laboratory investigations in secondary science. *Journal of Research in Science Teaching, 36*(10), 1065-1084.

Kluger, A. N., & Denisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*(2), 254-284.

Kluger, A. N., & Denisi, A. (1998). Feedback interventions: Toward the understanding of a double-edged sword. *Current Directions in Psychological Science, 7*(3), 67-72.

Krathwohl, D. R. (2002). A revision of bloom's taxonomy: An overview. *Theory Into Practice, 41*(4), 212-218.

Kuhn, D. (2005). *Education for thinking*. Cambridge, MA: Harvard University Press.

Kuhn, D. (2010a). Teaching and learning science as argument. *Science Education, 94*(5), 810-824.

Kuhn, D. (2010b). What is scientific thinking and how does it develop? In U. Goswami, (Ed.), *The Wiley-Blackwell handbook of childhood cognitive development*. Malden, MA: Wiley-Blackwell.

Kuhn, D., & Crowell, A. (2011). Dialogic argumentation as a vehicle for developing young adolescents' thinking. *Psychological Science, 22*(4), 545-552.

Kuhn, T. S. (1970). *The structure of scientific revolutions*. Chicago, IL: University of Chicago Press.

Lacey, A., & Wright, B. (2009). *Occupational employment projections to 2018*. Retrieved from www.bls.gov/opub/mlr/2009/11/art5full.pdf

Lane, D. (2008). *Free statistics book*. Retrieved May 1, 2015, from http://onlinestatbook.com/

Leach, J., & Scott, P. (2002). Designing and evaluating science teaching sequences: An approach drawing upon the concept of learning demand and a social constructivist perspective on learning. *Studies in Science Education, 38*(1), 115-142.

Leinhardt, G. (1983). Novice and expert knowledge of individual student's achievement. *Educational Psychologist, 18*(3), 165-179.

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

Lewis-Beck, M. S., Bryman, A., & Liao, T. F. (2004). *The Sage encyclopedia of social science research methods.* Thousand Oaks, CA: Sage.

Llosa, L. (2008). Building and supporting a validity argument for a standards-based classroom assessment of English proficiency based on teacher judgments. *Educational Measurement: Issues and Practice, 27*(3), 32-42.

Loucks-Horsley, S. (1998). *Designing professional development for teachers of science and mathematics*. Thousand Oaks, CA: Corwin Press.

Loughran, J., Berry, A., & Mulhall, P. (2012). *Understanding and developing science teachers' pedagogical content knowledge* (2nd ed.). Boston, MA: Sense.

Martin, M., Mullis, S., & Foy, P. (2012). *TIMSS 2011 International science report: findings from IEA's trends in international mathematics and science study at the eighth and fourth grades*. Chestnut Hill, MA: Boston College.

Marx, R. W., Blumenfeld, P. C., Krajcik, J. S., Fishman, B., Soloway, E., Geier, R., & Tal, R. T. (2004). Inquiry-based science in the middle grades: Assessment of learning in urban systemic reform. *Journal of Research in Science Teaching, 41*(10), 1063-1080.

McConnell, T. J., Parker, J. M., & Eberhardt, J. (2013). Assessing Teachers' Science Content Knowledge: A Strategy for Assessing Depth of Understanding. *Journal of Science Teacher Education*, 24(4), 717-743.

McComas, W. F. (1998). *The nature of science in science education: Rationales and strategies*. Dordrecht, The Netherlands: Kluwer Academic.

Meisels, S. J., Bickel, D. D., Nicholson, J., Xue, Y., & Atkins-Burnett, S. (2001). Trusting teachers' judgments: A validity study of a curriculum-embedded performance assessment in kindergarten to grade 3. *American Educational Research Journal, 38*(1), 73-95.

Mertler, C. A. (2002). *Using standardized test data to guide instruction and intervention.* College Park, MD: ERIC Clearinghouse on Assessment and Evaluation. (ERIC Document Reproduction Service No. ED470589).

Miles, J., & Shevlin, M. (2001). *Applying regression & correlation: A guide for students and researchers*. London, England: Sage Publications.

Morgan, A. (2006). *Feedback: Assessment for rather than of learning.* Retrieved June 14, 2006 from http://riel.bangor.ac.uk/the/documents/FEEDBACKJanuary06.ppt

Mulhall, P., Flowers, N., & Mertens, S. (2002). Understanding indicators related to academic performance. *Middle School Journal, 34*(2), 56-61.

Murphy, P. K., Wilkinson, I. A., Soter, A. O., Hennessey, M. N., & Alexander, J. F. (2009). Examining the effects of classroom discussion on students' comprehension of text: A meta-analysis. *Journal of Educational Psychology, 101*(3), 740-764.

National Research Council (2012). Inquiry and the National Science Education Standards. Washington, D.C.: National Academy Press.

Newton, P., Driver, R., & Osborne, J. (1999). The place of argumentation in the pedagogy of school science. *International Journal of Science Education, 21*(5), 553-576.

Next Generation Science Standards Lead States. (2013). *Next generation science standards: For states, by states.* Washington, DC: The National Academies Press.

Norris, S., L. Philips, & J. Osborne. (2007). *Scientific inquiry: The place of interpretation and argumentation*. In J. Luft, R.Bell, & J. Gess-Newsome *(*Eds.*), Science as inquiry in the secondary setting,* (73-75). Arlington, VA: NSTA Press.

Osborne, J., Erduran, S., & Simon, S. (2004). Enhancing the quality of argumentation in school science. *Journal of Research in Science Teaching, 41*(10), 994-1020.

Osborne, J., & Waters, E., (2002). Four assumptions of multiple regression that researchers should always test. *Practical Assessment, Research & Evaluation, 8*(2).

Packer, M. (2001). The problem of transfer, and the sociocultural critique of schooling. *Journal of the Learning Sciences, 10*(4), 493-514.

Park, S., & Chen, Y. (2012). Mapping out the integration of the components of pedagogical content knowledge (PCK): Examples from high school biology classrooms. *Journal of Research in Science Teaching, 49*(7), 922-941.

Parsons, S. A. (2012). Adaptive teaching in literacy instruction: Case studies of two teachers. *Journal of Literacy Research, 44*(2), 149-170.

Pedhazur, E. J. (2007). *Multiple regression in behavioral research.* United States: Academic Internet.

Perkins, D. N. (2009). *Making learning whole: How seven principles of teaching can transform education*. San Francisco, CA: Jossey-Bass.

Posner, G. J., Strike, K. A., Hewson, P. W., & Gertzog, W. A. (1982). Accommodation of a scientific conception: Toward a theory of conceptual change. *Science Education, 66*(2), 211-227.

Powell, K., & Kalina, C. (2009). Cognitive and social constructivism: Developing tools for an effective classroom. *Education, 2*(13), 241-250.

Reigeluth, C. M. (1999). *Instructional design theories and models: A new paradigm of instructional theory*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Sadler, T. D. (2006). Promoting discourse and argumentation in science teacher education. *Journal of Science Teacher Education, 17*(4), 323-346.

Sandoval, W. A., & Reiser, B. J. (2004). Explanation-driven inquiry: Integrating conceptual and epistemic scaffolds for scientific inquiry. *Science Education, 88*(3), 345-372.

Sawada, D., Piburn, M. D., Judson, E., Turley, J., Falconer, K., Benford, R., & Bloom, I. (2002). Measuring Reform Practices in Science and Mathematics Classrooms: The Reformed Teaching Observation Protocol. *School Science and Mathematics, 102*(6), 245-253.

Sawyer, R. K. (2006). *The Cambridge handbook of the learning sciences*. Cambridge, MA: Cambridge University Press.

Scott, P. (2008). Teacher talk and meaning making in science classrooms: A Vygotskian analysis and review. *Studies in Science Education, 32*(1), 45-80.

Scott, P. H., Mortimer, E. F., & Aguiar, O. G. (2006). The tension between authoritative and dialogic discourse: A fundamental characteristic of meaning making interactions in science lessons. *Sci. Ed. Science Education, 90*(4), 605-631.

Scruggs, T. E., Mastropieri, M. A., Bakken, J. P., & Brigham, F. J. (1993). Reading versus doing: The relative effects of textbook-based and inquiry-oriented approaches to science learning in special education classrooms. *The Journal of Special Education, 27*(1), 1-15.

Sharpley, C. F., & Edgar, E. (1986). Teachers' ratings vs standardized tests: An empirical investigation of agreement between two indices of achievement. *Psychology in the Schools, 23*(1), 106-111.

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher, 15*(2), 4-14.

Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review, 57*(1), 1-23.

Shulman, L. S., & Quinlan, K. M. (1996) The comparative psychology of school subjects. In D. C. Berliner & R. C. Calfee (Eds.), *Handbook of educational psychology*. (399-422). New York: Simon & Schuster Macmillan,.

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

Siddiquee, S., & Ikeda, H. (2013). Science talk in the secondary classrooms: Analysis of teachers' feedback. *European Scientific Journal, 9*(11), 66-71.

State of the Union Address. (2011). Retrieved January 20, 2015, from https://www.whitehouse.gov/the-press-office/2011/02/12/remarks-president-state-union-address.

Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology, 104*(3), 743-762.

Taylor, J. C., Therrien, W. J., Kaldenberg, E., Watt, S., Chanlen, N., & Hand, B. (2011). Using an inquiry-based teaching approach to improve science outcomes for students with disabilities: Snapshot and longitudinal data. *Journal of Science for Students with Disabilities, 15*(1), 1-14.

Tobin, K. G. (1993). *The Practice of constructivism in science education*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Tobin, K., & Llena, R. (2010). Producing and maintaining culturally adaptive teaching and learning of science in urban schools. *Cultural Studies of Science Education, 1*, 79-103.

Vogt, F., & Rogalla, M. (2009). Developing adaptive teaching competency through coaching. *Teaching and Teacher Education, 25*(8), 1051-1060.

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

Wallace, C. S., Hand, B., & Prain, V. (2004). *Writing and learning in the science classroom*. Dordrecht, The Netherlands: Kluwer Academic.

Weiss, R, Pasley, J, Smith, S, Banilower, R, & Heck, J (2003). *Looking Inside the Classroom: A Study of K–12 Mathematics and Science Education in the United States.* Chapel Hill, NC: Horizon Research.

West, E. A., Paul, C. A., Webb, D., & Potter, W. H. (2013). Variation of instructor-student interactions in an introductory interactive physics course. *Physical Review Special Topics - Physics Education Research, 9*(1).

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.

Winne, P. H., & Butler, D. L. (1994). Student cognition in learning from teaching. In T. Husen & T. Postlewaite (Eds.),*International encyclopedia of education* (2nd ed.., pp. 5738–5745). Oxford, UK: Pergamon.

Wright, D., & Wiese, M. (1988). Teacher judgment in student evaluation: A comparison of grading methods. *Journal of Educational Research, 82*, 10-14.

Zee, E. H., & Minstrell, J. (1997). Reflective discourse: Developing shared understandings in a physics classroom. *International Journal of Science Education, 19*(2), 209-228.

APPENDIX A

TEACHER FEEDBACK OBSERVATION TOOL

Dialogic Teaching Characteristics Dialogic: Teacher and students consider a range of ideas. If the level of discourse is high, they pose genuine questions as they explore and work on different points of view. If the level of discourse is low, the different ideas are simply made available.														
A-Development of the views shared by the scientific community. Shifting from an everyday view to a scientific view.			B-Facilitating Dialogue			C-Providing Opportunities for reflection			D-Establishing classroom environment.			E-Developing understanding		
A1	A-1	A-3	B-1	B-2	B-3	C-1	C-2	C-3	D-1	D-2	D-3	E-1	E-2	E-3

A1 - Asks students if their ideas fit under the “Big Idea”

A2- Asks students to compare their views with the views of the scientific community

A3- Asks student for clarification of their claim

B-1- Asks students to directly address another student. (“Can you answer her question?”

B-2- Asks students to make connections between more than one discussion point (Example: How do you think energy and force are related?)

B-3- Focuses the conversation on a topic/question/claim raised by another student.

C-1- Asks student to reflect on their views (have they changed?) written or oral.

C-2- Asks students to make consensus making statements based on peer dialogue. (Do you agree with what your classmate said?)

C-3- Asks students if they agree with the statements of expert (The book says _____, can you show me evidence of this?)

D-1- Asks students to recognize alternative points of view.

D-2- Teacher asks students to provide evidence that supports their claims

D-3- Teacher does not correct the view of a student even when it does not agree with the view of the scientific community.

E-1- Teacher asks students to raise questions or explain observations

E-2- Asks students for personal interpretation.

E-3- Asks student to use their schema to make a hypothesis or share ideas before any science content is shared.

APPENDIX B

EXAMPLE OF BASIC-CONTENT LEVEL QUESTION

What is the formula for density?

a) Density = buoyancy / weight

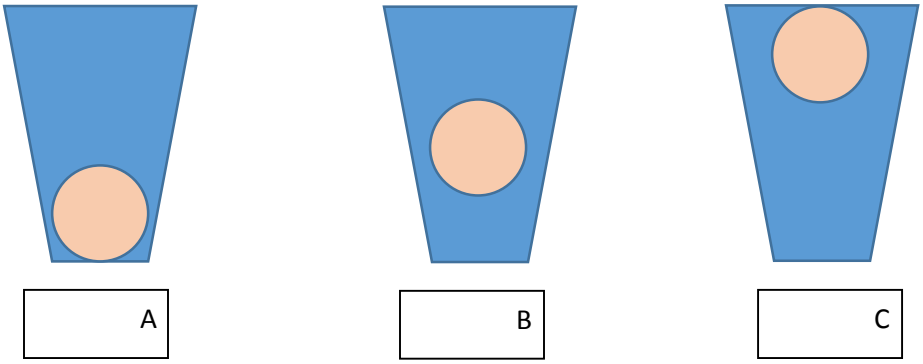
b) Density = mass / volume

c) Density = buoyancy / mass

d) Density = weight / volume

APPENDIX C

EXAMPLE OF PHENOMENOLOGICAL LEVEL QUESTION



In the experiment above all three balls have the same density. What can you tell me about the density of the three liquids?

A. Liquid A is less than liquid B
B. Liquid C is greater than liquid B
C. Liquid B is less than liquid A
D. Liquid C is greater than liquid A

APPENDIX D

EXAMPLE OF RULE-BASED QUESTION

What do hot air in the atmosphere and hallow Ping-Pong balls in a bucket of water have in common?

A. Both have a white color
B. Both move fast because of they have low density
C. Both have densities that are higher than their medium.
D. Both rise to the top of their medium because of their low density

APPENDIX E

SAMPLE OF TEACHER HIT-RATE JA FORM

Student 1	
Problem 1	Correct Y or N
Problem 2	Correct Y or N
Problem 3	Correct Y or N
Problem 4	Correct Y or N
Problem 5	Correct Y or N
Problem 6	Correct Y or N
Problem 7	Correct Y or N
Problem 8	Correct Y or N
Problem 9	Correct Y or N

APPENDIX F

TOLERANCE, VIP, SKEW, AND KURTOSIS MEASUREMENTS

	Tol	VIF
JA	.974	1.026
DF	.878	1.139

	Skew	Kurtosis
JA	-0.01	-1.13
DF	0.29	-1.06